# A Machine Learning Approach to Crystal Structure Prediction

by

Christopher Carl Fischer

B.S. Metallurgical and Materials Engineering
Colorado School of Mines (2002)

Submitted to the Department of Materials Science and Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2007



Author .....
Department of Materials Science and Engineering
August 16, 2007

Certified by ..... 8/16/07
Gerbrand Ceder
R.P. Simmons Professor of Materials Science
Thesis Supervisor

Accepted by .....
Samuel M. Allen
POSCO Professor of Physical Metallurgy
Chair, Departmental Committee on Graduate Students

# A Machine Learning Approach to Crystal Structure Prediction

by

Christopher Carl Fischer

Submitted to the Department of Materials Science and Engineering
on August 16, 2007, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

## Abstract

This thesis develops a machine learning framework for predicting crystal structure and applies it to binary metallic alloys. As computational materials science turns a promising eye towards design, routine encounters with chemistries and compositions lacking experimental information will demand a practical solution to structure prediction. We review the ingredients needed to solve this problem and focus on structure search. This thesis develops and argues for a search strategy utilizing a combination of machine learning and modern quantum mechanical methods.

Structure correlations in a binary alloy database are extracted using probabilistic graphical models. Specific correlations are shown to reflect well-known structure stabilizing mechanisms. Two probabilistic models are investigated to represent correlation: an undirected graphical model known as a cumulant expansion, and a mixture model. The cumulant expansion is used to efficiently guide Density Functional Theory predictions of compounds in the Ag-Mg, Au-Zr, and Li-Pt alloy systems. Cross-validated predictions of compounds present in 1335 binary alloys are used to demonstrate predictive ability over a wide range of chemistries – providing both efficiency and confidence to the search problem. Inconsistencies present in the cumulant expansion are analyzed, and a formal correction is developed.

Finally, a probabilistic mixture model is investigated as a means to represent correlation in a compact way. The mixture model leads to a significant reduction in model complexity while maintaining a level of prediction performance comparable to the cumulant expansion. Further analysis of the mixture model is performed in the context of classification. Unsupervised learning of alloy classes or groups is shown to reflect clear chemical trends.

Thesis Supervisor: Gerbrand Ceder
Title: R.P. Simmons Professor of Materials Science

*To my parents, David and Jeretta*

# Acknowledgments

This thesis was not created out of a vacuum, rather it grew out of a rich environment deserving most of the credit. First, I must thank my advisor, Gerd Ceder for fueling this thesis work with plenty of enthusiasm, funding, computing power, and most importantly, ideas. Gerd is a master of communicating research results to his audience and it has been a pleasure to watch a healthy dose of flair combine with pedagogy to entertain and teach the most thorny scientist. I would also like to thank my committee: Sam Allen and Nicola Marzari for their helpful suggestions and unique insight, and Francesco Stellacci for his help during the final hour. The Department of Energy and Institute of Soldier Nanotechnologies are acknowledged for funding. In addition, I'm grateful to the Presidential Fellowship program at MIT for taking the stress out of first-year funding and the Department of Materials Science and Engineering for Teaching Assistantships during the Fall of 2005 and 2006.

The Ceder research group has been a constant source of support and inspiration for their talent, creativity, and never-ending juicy stories. In particular, I must thank: Eric (`thewu.mit.edu`) Wu for his time and encouragement during my first year as a graduate student – Wu taught me how to run experiments on the computer, grease the wheels, and keep priorities in line; Dane (the dog) Morgan, Professor Morgan that is, maintained his generosity and patience as a Postdoctoral Researcher despite being stuck in a tiny office with a first-year graduate student; Anton Van der Ven (Professor that is) for teaching me statistical mechanics; Tim (t-bone) Mueller for pointing me in the right direction when I needed it, especially when I didn't know I needed it; and Kristin Persson for her technical insight, sense of humor, and companionship throughout the good and bad. The work of Stefano Curtarolo, Kristin, Dane, and Gerd provided a significant milestone needed to pursue the work in this thesis and the foundation for moving forward. Kevin Tibbetts provided all of the (cleaned) Pauling File data used throughout my research – without it, this work would have never happened. Over the years discussions with Fei, Matteo, Kisuk, Caetano, Byungchan,

Shirley, and Maria have taught me many things. I can't wait to see what is in store for Byoungwoo, Lei, Yoyo, Xiaohua, Geoffroy, Charles, Anubhav, and Shyue.

My roommates Carl (c-monstah) Dohrman, Asher (the dude) Sinensky, Scott Litzelman, and Tim, have defined home, and kept it engaging for four years. Carl's baked goods, Tim's cookies and enviable recliner, Scott's prolific knowledge of the silver screen, his planning capabilities, and Asher's ability to test those plans will never be forgotten. My friends here in Boston balanced out the MIT haze: Professor Megan Frary, Bobby Boyer, Andy Albers, Bob Barsotti, Christina Kaba (and of course Jack), Elisa Alonso, Rachel Pytel, Tiffany Santos, Gretchen DeVries, Andrew Detor, Jay Trelewicz, John Mills, and the always lively David Danielson. Jay, Eric, Zach, always men of principle, have provided crucial support from afar.

My parents have always been committed to cultivating my education no matter which direction I decided to go. Together, they have invested more care and time preparing this thesis than anyone.

Most of all, MIT led me to Sharon, who provides a happiness I didn't know was possible. She has made life wonderful for the last four years, especially by maintaining my sanity during the process of "graduating soon".

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

*The primary goal of this thesis is to develop a framework in which historical data is used in conjunction with modern quantum mechanical methods to predict the crystal structure of a material.*

Calculating material properties through the use of so-called first principles methods is transforming fundamental materials science. Owing to steady improvements in both computing technology and the availability of robust, thoroughly tested *ab initio* total energy codes, the properties of materials may be rapidly predicted, from scratch, in a *reliable* manner [5, 3, 6, 7, 8]. These advances are the enabling agents in a virtual materials design laboratory whereby materials with targeted properties are understood and optimized *prior to* synthesis. This new process of discovery spans the full range of time and length scales with applications in virtually all areas of materials science. Current applications include materials for energy conversion and storage (e.g., battery cathodes [9, 10, 11, 12] and anodes [13], hydrogen storage [14, 15, 16, 17, 18], fuel cell cathodes [19], thermoelectrics, optimized metallic alloys for stability [20] or mechanical behavior, nanotubes for electronic devices [21] and high-rate batteries, dielectric resonators [22]. This list is far from complete, but gives an indication of the breadth of impact that first principles techniques are making in the search for new materials. Recognizing this impact, a recent issue of the Bulletin of the Materials Research Society focused on the many material properties that can now be addressed in a fully *ab initio* manner [5]. To put it bluntly, the first principles search is on, and

the stage set for these techniques to venture into new chemical territories, designing materials before they are synthesized and tested.

To this day, first principles calculations have focused on understanding and optimizing the behavior of *known* materials with atomistic level detail. In such a framework experimental structural information is used as a starting point for detailed first principles calculations. Combining the detailed information and control afforded by DFT calculations with experiment has led to a much richer understanding of known materials and the causal mechanisms leading to their observed behavior. However, when exploring truly new chemistries the most basic information required to run a calculation, namely the positions of atoms in a crystal, is missing. We call this the "structure prediction" problem [23] and despite significant advances that are highlighted throughout this chapter, predicting structure remains an elusive, difficult task. This thesis is devoted to developing a strategy for solving the structure prediction problem through a combined use of historical information and first principles calculations. The next section details the problem of predicting structure, the ingredients required to solve it, and summarizes the literature on the subject.

## 1.1 the structure problem

Crystal structure occupies both a fundamental and widely applicable role in materials science. Many relevant physical properties of inorganic materials are directly tied to, and sometimes prohibited by, the underlying symmetry of their crystalline form [24]. Structure alone has a pronounced effect on a wide range of properties from band gaps to brittle fracture, so material property calculations *ab initio* quickly lose their relevance and impact when performed on the *wrong* structure. Therefore, to fully harness the capabilities of first principles methods we require a strategy to predict structure, i.e., given a material's composition and a set of thermodynamic control variables, what is the stable state of the system ? Methods for predicting structure generally require three ingredients: an accurate **energy** model, an effective strategy to **search** through the space of possible structures, and a method for evaluating

**entropy.**

At the energy and time scales of interest to the material's scientist, well established methods exist for evaluating the energy of a system. To do so, one must solve the Schrodinger equation for a collection of interacting electrons and nuclei. For most materials, the electronic degrees of freedom are assumed to respond instantaneously to an electrostatic potential created by a clamped set of nuclei (i.e., the Born-Oppenheimer approximation)[1]. The quantum mechanical problem of computing the energy of a set of electrons under an applied external potential has been successfully addressed through well known approximations to Density Functional Theory (DFT) [26, 27, 28, 29] whereby the intractable many-body problem is mapped onto an effective single particle theory that is *exact* in principle. It can be argued that practical implementations of DFT, such as the Local Density Approximation or LDA, employ somewhat uncontrolled approximations. However, because DFT has proved itself so immensely effective in reproducing a wide range of experimental properties [5], predictions with DFT are almost treated as gospel. Most importantly for structure prediction, where one is interested in determining the stable state of a system, DFT has been shown to correctly reproduce energy *differences* between crystal structures [6]. For example, Curtarolo, Morgan, and Ceder[3] found that for metals and their alloys, DFT in the LDA or Generalized Gradient Approximation (GGA) reproduced the true ground state of a system with substantial success ($90-97\%$). There are some classes of systems, such as transition metal or mixed valence oxides, where LDA and GGA are known to yield results that are in *qualitative* error. These failures can often be traced back to the self interaction error present in DFT which tends to delocalize states that should be localized [30]. Pronounced errors are manifested in electron transfer reactions between delocalized and localized orbitals, breaking the usual DFT "cancellation of errors" and measurable quantities such as redox potentials are in significant error [11]. For some chemistries the errors are more severe, and DFT might predict compound formation while nature phase separates. To circumvent this issue,

[1]though notable exceptions exist where the zero-point motion of the nuclei are large (superfluids), systems where large electron-phonon interactions are present (superconductors), and other non-adiabatic systems [25]

the DFT+U approach and more elaborate methods such as dynamical mean field theory (DMFT) have been shown to improve agreement between theory and experiment [11, 31]. In summary, some highly celebrated errors of DFT remain, but history has shown that DFT (or its close relatives) will assign the correct energetic *order* over structures with a high degree of accuracy.

While an accurate energy model is necessary to predict structure, it alone is not sufficient. In general, one will be interested in the stable state of a system held at finite temperature. Under these circumstances the system will fluctuate over a great number of states consistent with the imposed thermodynamic boundary conditions. These fluctuations imply that the state of the system cannot be described by a single microstate, but can only be understood in terms of an appropriate statistical average over all allowed microstates. Thus, once the ground state of a system has been found, methods for evaluating entropy require knowledge of the material's excitation spectrum detailing both the state space or type of excitations (e.g. vibrational, configurational, electronic), and the energetics of those excitations [32]. For many years, models of excitations in solids have been studied extensively (refs) and the dominant mechanisms through which a system *equilibrates* with its environment are well known. Once an appropriate excitation spectrum has been constructed, a strategy is needed for appropriately evaluating thermal averages. A variety of methods such as Molecular Dynamics [33], Monte Carlo sampling [34], mean field theories [35, 36, 37], and occasionally analytic solutions exist to evaluate a system's entropy.

While challenges certainly remain in all three ingredients required for structure prediction, this thesis focuses on a key missing piece, namely an efficient method for *searching* through the space of possible ground state structures. We confine our problem to searching for the ground states of a system as a function of composition at zero temperature and pressure as this is a key ingredient (boundary condition) for the structure problem under other conditions of equilibrium. The next section outlines the search problem and strategies for its solution.

## 1.2 Ground state structure search

Searching for ground states is a difficult, highly non-linear optimization problem of a system's energy in the space of its atomic coordinates[2]. Throughout the course of history the methods of searching for ground states tend to fall into two categories: (1) those which directly optimize a system's atomic coordinates and (2) heuristic rules governing structural stability. Splitting the world of structure search into these two camps and understanding the advantages and limitations of each, provides the underlying context for the solution presented in this thesis.

### 1.2.1 Coordinate based search

Optimizing the energy of a system in the space of atomic coordinates and unit cell parameters is made tremendously difficult due to the high dimensionality of the space coupled with the presence of many local energy minima. Canned optimization [38] of a system's energy converge to one of the many local minima and to address this issue, several strategies have been developed. One method of reducing this complexity is to optimize a system's energy starting from a small number of common **structure prototypes** because such prototypes are known to be the global minimum configuration in *some* chemistry. This approach rests on the observation that in nature, a relatively small number of structure prototypes are observed across a large number of chemical systems (e.g., the CsCl or NaCl structure prototypes have been observed in $\approx 7\%$ of all binary systems studied experimentally). Researchers select a finite number of prototypes based on their own intuition, a highly biased process in itself, perform a gradient based optimization on these, and proclaim the resulting stable structure as the true ground state. Unfortunately, there are currently more than 2500 known structure prototypes. A truly unbiased strategy would require calculating the energy of each known prototype for every chemistry, which is simply an intractable solution. Due to the presence of many local minima strategies have

[2]For a periodic system consisting of N particles there are $6 + 3(N - 1)$ degrees of freedom. If one makes an assumption about the symmetry of the system this number may be reduced.

been developed which essentially "hop" out of the local optimum. These alternative methods optimize a system's coordinates with a stochastic optimization technique such as Simulated Annealing (SA) [39] or a Genetic Algorithm (GA) [40]. A key advantage of stochastic optimization methods is that they can place a guarantee on finding the global optimum, albeit in the "infinite runtime" limit. In SA standard gradient-based optimization is paired with a source of random perturbation e.g., by coupling the system to a fictitious heat bath held at constant temperature. By gradually lowering the temperature of the heat bath the system will eventually "cool" into the ground state configuration. In practice [41] one finds that SA is both slow to converge and will often "miss" energy minima even for moderate cooling rates. The GA approach [42, 43, 44] optimizes a system's coordinates over a population of structures. Generations of structures are evolved by mating fit members of the population (e.g., by combining slices of two energetically favorable structures) subject to mutation (random perturbations of a population member's coordinates). The key shortcoming of stochastic optimization techniques is the significant number of energy evaluations required for accurate convergence. For example, in their use of a GA Probert, et al. [41] and Trimarchi, et al. [45] required more than 40 and 50 local optimizations respectively to obtain the structure of bulk silicon. In multicomponent systems the computational cost of GA is increased due to additional configurational degrees of freedom (atom swaps). A study by Glass, et. al. [44] required 390 total energy relaxations to find the structure of $MgSiO_2$, and Trimarchi, et. al required 70 to obtain the structure of GaAs. In metallic alloys, energy differences between structures are more subtle than "octet" compounds and energy excitations required to create 0-,1-,and 2-dimensional defects are quite small. Thus, finding ground states in metallic alloys is likely to be the most stringent test for a GA (e.g. in Reference [45] the authors simply stopped after 40 calculations having a structure 2meV/atom above the ground state for $Au_8Pd_4$). Note that because the size of the primitive cell is unknown *a priori*, a supercell must be employed to allow sufficient flexibility for the GA, adding a significant overhead in computational cost.

An alternative approach to optimizing a system's coordinates is to restrict the sys-

tem to a *fixed topology*. Such methods have been tremendously useful for chemistries where the set of phases appearing as a function of composition can be described as an **ordering** or decoration on a single or small set of underlying parent topologies, e.g mixing atoms "A" and "B" on an fcc or bcc lattice. By constraining the description of a system to a fixed topology, the standard optimization problem involving real-valued coordinates of all atoms and unit cell parameters can be converted to an integer programming problem [46, 47, 48] consisting of finding the lowest energy decoration of the topology. In this scheme, the energy of a system can be rigorously expanded with respect to the occupation of groups of nearby lattice sites resulting in a so-called Cluster Expansion (CE) [49, 35, 36, 50].

$$\begin{aligned} H(\sigma_1, \sigma_2, \ldots, \sigma_N) &= V_0 + \sum_i V_i \Phi_i(\sigma_i) + \frac{1}{2} \sum_{i,j} V_{i,j} \Phi_{i,j}(\sigma_i, \sigma_j) + \cdots \\ H(\vec{\sigma}) &= \sum_\alpha V_\alpha \Phi_\alpha(\vec{\sigma}) \end{aligned}$$

Here $\sigma_i$ is a variable denoting the species occupying site $i$ and $\Phi_\alpha(\vec{\sigma})$ is an orthogonal polynomial of occupation variables in the cluster $\{i \in \alpha\}$ and $V_\alpha$ is an expansion coefficient. In practice, a CE is truncated after including a small number of terms such as point, nearby pairs and triplets, and so on. Remarkably, the ground states for some truncated CE's can be obtained exactly [46, 47, 35, 37]. More frequently, to accurately describe the energetics of mixing, a CE contains so many terms that only approximate solutions can be made [51, 48]. Often the $V_\alpha$ are fit [52, 53] to a set of energies obtained via first principles calculations, and the ground states are determined for a fixed set of interaction coefficients. Such an approach has led to several structure predictions [9, 54, 8] some of which have later been confirmed by experiment [55, 56]. The CE has recently been "extended" to include the configurational energetics of localized electrons [57], mixtures of cations and anions [58], as well as rotational degrees of freedom of fixed structural units such as OH or NH groups [15]. Zhou et al. recently used the CE to coarse grain protein energetics [59], in essence 20-component alloy problem ! These extensions provide continued evidence of the utility of such coarse-grained [50] or "restricted degree of freedom" Hamiltonians. However, because

the CE restricts a system to a fixed topology, solving all structure prediction problems with the CE is simply not a tractable solution. In many chemistries, the set of phases appearing at different compositions span many different topologies, and performing the full procedure of constructing a CE for each is simply too computationally costly.

At their core, each method mentioned in this section is essentially a mathematical optimization technique coupled to varying levels of approximation to a system's energy. As such, all suffer from what I call a lack of knowledge transfer across chemistries. For example, using any of the above techniques to exhaustively determine the ground states of the Au-Sc system gives no indication of the likely ground states in the Au-Zn system. Ultimately this task of transferring knowledge from system to system has been left to the researcher, utilizing his or her own chemical intuition to decide *what* to calculate. In the context of deciding what to calculate, simple heuristic guidelines, discussed next, are often used to winnow the enormous set of candidates into one that is both chemically reasonable and computationally tractable.

### 1.2.2 Heuristic methods

Historically, structural stability has been understood in a qualitative manner using heuristics: simple, efficient rules and guidelines appropriate for classes of chemistries which work well under most circumstances at the expense of introducing bias. These guidelines are used for such tasks as rationalizing why some structures are stable over others, or deciding if two elements will mix with a positive or negative enthalpy of mixing $\Delta H_{\text{mix}} \lesssim 0$. Heuristics don't rely on any microscopic Hamiltonian, but rather attempt to relate stability to simple physical concepts. Perhaps the most widely known set of heuristic rules for understanding structure stability, the so-called Pauling Rules, were formulated in 1929 by Linus Pauling [60] and detailed later in his book *The Nature of the Chemical Bond* [61]. The Pauling Rules are a set of intuitive rules governing both the type and connectivity of the basic structural units making up *ionic* compounds. They owe much of their success to treating both geometrical (or space filling) and chemical concepts on the same footing. Similar characterizations

based on structural units have also been put forth for metals.

**Size effects and hard sphere packing in metals** Due to the delocalized nature of valence electrons in metals, crystal structures of metallic alloys are often conceptualized as dense packings of neutral hard spheres. While at first glance this approximation may seem too severe, the existence of several *classes* of metallic alloy structures were succinctly summarized on such a "size-effect" basis by Laves and Witte [62] in 1935. According to their reasoning, when two elements of very different size are mixed (tacitly assuming that the constituents would otherwise prefer to mix $\Delta H_{mix} < 0$), some structures are preferred over others due to a favorable accommodation of the size mismatched atoms. Following on these ideas, Frank and Kasper [63] put forth a geometric theory of topologically close-packing phases, providing a rational for the stability of a number of structures such as the $\sigma$, $\mu$, $\beta$-tungsten, and Laves phases. Somewhat recently, Daams, et al. [64] has identified a small number of structural units or local environments recurring across a large population of metallic compounds. Size effect and packing arguments are a useful tool for *characterizing* a particular structure, but translating these characterizations into a *predictive* model of stability is somewhat difficult and has been met with only limited success. Furthermore, understanding crystal structure on the basis of their local environments alone carries an inherent limitation in descriptive power. Truly inequivalent structures will appear identical on the basis of their local environments, but differ in how the environments are fit together.

**Electronic and chemical arguments** Early understanding of the physics of metals, before the days of DFT, pseudopotentials, and GigaFLOP commodity computers relied on simplified descriptions of how a materials constituents interacted. The analyses performed were borne out of the solid state physics community who identified several key factors influencing compound formation in alloys.

Hume-Rothery noted in 1926 [65, 66] that the ratio of "valence" electrons to atoms in a compound appears constant for some intermetallic phases (e.g., CuZn and

$Cu_5Sn$ both have a ratio of $e/a = 3/2$). Mott and Jones [67, 68], using ideas put forth by Bloch, deduced a formal relationship between these "magic" ratios and structure stability. By combining a rigid band theory of electrons with valence electron concentration, Jones was able to illustrate how the nearly free electron Fermi surface would encounter a zone boundary (and hence a divergence in the electronic density of states) providing an energetic preference for one structure over the other. While the severity of approximations in Jones' theoretical analysis are now more well understood, the idea that the valence electron concentration plays an important role in the formation of compounds persists to modern treatments of the same problem [69]. Perhaps the most successful heuristic model of compound formation is due to Miedema [70, 71], which *classifies* binary alloy systems as "compound forming" or "phase separating" based on a two-parameter model for $\Delta H_{mix}$. Although the Miedema model cannot discriminate between structures, it has been found to be roughly 95% correct in predicting whether a system will form compounds or not.

**Structure Maps** One of the key concepts to arise out of heuristic methods is the idea that structural stability can be understood on the basis of simple, physically motivated parameters such as size mismatch ($\Delta r$), valence electron concentration ($n_e$), electronegativity difference ($\Delta\chi$), etc. These parameters provide a simplified and compact indication of the mixing tendencies of a materials constituents. A *structure map* is constructed by plotting which structures are stable over a wide range of systems in a coordinate space defined by these heuristic parameters. When the coordinates dictating structural stability, such as $\Delta r$ and $\Delta\chi$, are similar in value for different chemistries it seems reasonable to expect the stable structures in both to be the same or very similar. Figure 1-1 gives a schematic representation of an idealized structure map for several AB compounds. The coordinates $x_1$ and $x_2$ could be $\Delta r$ and $\Delta\chi$ or some other set of physically relevant parameters. The ideas behind structure maps originated in the work of Mooser and Pearson [72] using the coordinates $\Delta\chi$, a measure of ionicity, and average principle quantum number, which they took as a measure of the directionality of bonding. Subsequently Harrison, Heine, Simons,

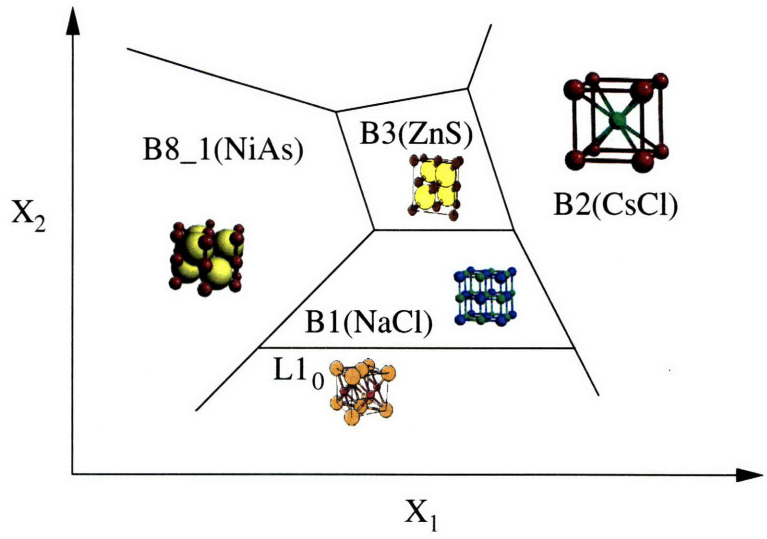


Figure 1-1: Idealized structure map of AB compounds. Each structure is stable in regions of the parameter space defined by the coordinates $x_1$ and $x_2$ which could be size mismatch, electronegativity difference, valence electron concentration, etc..

and St.John [73, 74, 75, 76, 77] identified structural trends using parameters derived from pseudopotentials. Zunger [78] later used these concepts in his own attempt to "systematize" binary compounds. Phillips and Van Vechten [79] proposed a dielectric classification of the crystal structures of octet compounds, making a point to connect the classification task to experimentally measured quantities. Villars [80, 81, 82] created a wide number of structure maps using various combinations of the above parameters even creating "property" maps – the equivalent of a stability map, but for material properties. Structure maps for *oxides* have also been constructed by Muller and Roy [83].

The goal of structure mapping is to achieve separation between different structures or at least different structure classes and as such it can be viewed as a *classification* problem [84] – i.e., in Figure 1-1 all of the B2-forming systems are one class, while those forming the B3 structure are another. In other words, chemical systems are *classified* by their parameter values in terms of the stable structures that appear. Given an objective of classification, many heuristic coordinate combinations are known to fail [4] (i.e., multiple structural domains significantly overlap). For example, using the coordinates $\Delta r$ and $\Delta\chi$ alone will mix structures forming between sp-bonded

materials and transition metal intermetallics. To overcome this deficiency, additional coordinates such as total electron count must be introduced, a strategy employed by Villars [80]. Worse still, common combinations of coordinates such as $\Delta r$ and $\Delta\chi$ cannot be varied independently [85]; for example, the electronegativity of an element can be loosely related to many measures of atomic "size". The key point is that one particular set of coordinates may work well for covalently bonded materials, but fail for metallic or non-directionally bonded materials. Pettifor, recognizing the deficiencies in using such coordinates created his own chemical scale, $\chi$, indexing each element to a unique number. This chemical scale was then used to create Pettifor's version of structure maps for common stoichiometries in binary alloys [4]. Pettifor's chemical scale, while still physically motivated (tending to run up and down the columns of the periodic table), achieves better separability (classification) at the expense of understanding the underlying mechanisms driving stability.

**Heuristics summary** Understanding structural stability on the basis of heuristic arguments is a powerful method of efficiently summarizing structural trends **across** a large number of chemistries. In essence, these methods forgo atomistic level detail in an attempt to extract general trends. As such they provide just the transfer of knowledge that is notably absent in coordinate-based searching for stable structures.

Although heuristic methods provide a powerful approach to extracting general structural trends, a number of thorny issues beyond those already mentioned remain. First, while heuristics might identify some favorable structures for a given chemistry, they provide no explicit ranking. Thus structures are either likely or unlikely, but the question still remains, "by how much ?" Perhaps the best strategy one can use is a ranking of candidate structures by some function of distance to the system of interest in the structure map [86]. A more severe limitation is the lack of a strategy for dealing with conflicting information. For example, an experimental database may contain information about multiple structures appearing at the same composition, presenting a complication when constructing a structure map: which structure should be used in the map ? Given that both experiments appear valid, and barring additional

discriminating information, both structures must be considered a valid assignment. To utilize a structure map a decision must be made, albeit at the expense of removing a considerable amount of data from the dataset (e.g. Morgan et al. [86] removed 40% of the available data based on this complication). Finally, structure maps are constructed for each composition independently, although it seems quite intuitive that correlation should extend from one composition to another. Because of this limitation, information available in one structure map cannot be utilized to perform predictions at others.

## 1.3 Our solution

We believe significant progress towards the structure prediction problem can be made by combining the suggestive character of crystal structure correlations present in historical data with the accuracy of modern first principles energy methods. In a sense, this thesis will attempt to combine the advantages of both the heuristic and coordinate-search techniques outlined in this chapter while avoiding the pitfalls associated with both. Figure 1-2 shows an outline of our overall algorithm. Structural correlations are extracted from a database of experimental and/or computed structure information. These correlations are in turn used to suggest the most likely candidate structures for a given chemistry, and the stability of the candidates is assessed with an accurate energy model based on modern implementations of DFT. Stable structures are fed back into the database, correlations rebuilt, and additional iterations are performed as needed. Provided a sufficient machine learning method can be found to extract and utilize structure correlations, this framework will solve both the *knowledge transfer* problem associated with coordinate-based optimization schemes and the *accuracy* problem associated with heuristic rules.

The primary focus of this thesis is therefore the construction of a general, unbiased technique for extracting and using structure correlations present in historical information. Of the three ingredients needed for predicting structure, accurate models for both the energy and entropy exist for all chemistries and structures considered

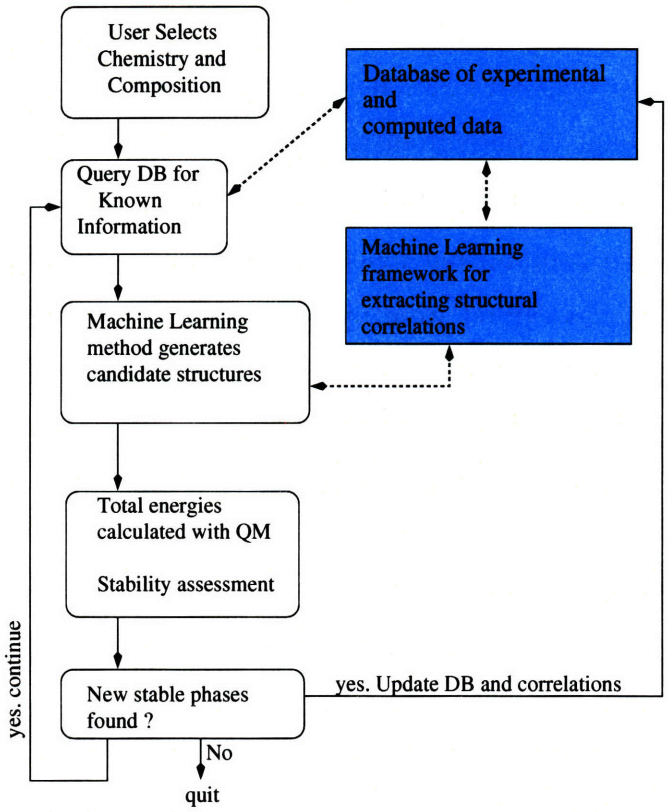


Figure 1-2: Outline of formalism for determining the ground states for a specified chemistry and composition.

within this thesis. The machine learning techniques developed here will ultimately have a significant impact on how one searches through the space of crystal structures based on prior information. Our formalism will be described in Chapter 2 and specific implementations of the technique will be described in Chapters 3 and 4. We have used our technique to efficiently determine the ground states in the Ag-Mg, Au-Zr, and Li-Pt alloy systems, the results of which are summarized in Chapters 3. Chapter 5 gives some concluding remarks and an outlook for future work.

# Chapter 2

# Probabilistic models of structure correlation

*Chapter 1 reviewed the structure prediction problem, the ingredients needed to solve it, and a brief history of available literature. This chapter introduces the conceptual framework used in the remainder of the thesis.*

To predict the stable state of a system under a particular set of thermodynamic boundary conditions we require accurate models of the energy and entropy of a system, and an effective procedure for searching through the space of structures. Our strategy is to develop a machine learning technique to enable a systematic, informed search over the space of structures using a modern DFT-based Hamiltonian. It should be noted that if we can narrow the list of possible structures down to just a handful of most likely candidates, the stability of these candidates can be accurately and rapidly evaluated with DFT. In essence, by systematically winnowing the set of candidates, we are solving the structure prediction problem within the class of *known* possibilities. In this chapter we will use a machine learning technique to extract correlations from historical data (in essence a form of compression) with an eye towards remaining unbiased. These correlations will subsequently be used in later chapters for *prediction.*

## 2.1 Data Abstraction

If one is to use any machine learning method to operate on data, perhaps the first step of the process is to define what your variables are, and give some indication of the values that they can take on. This process of abstracting the raw data into a form that is amenable for calculation is of critical importance at all stages down the machine learning production line. How the data is represented will determine how difficult it will be to extract and utilize correlation within the data. We are interested in obtaining correlations from raw experimental and computed information on structural stability. We are of course at liberty to choose any set of variables to describe structural correlation, some examples of which have been presented in our discussion of heuristic coordinates in Section 1.2.2. However, the serious deficiencies of such a set of coordinates are well-known. Therefore, we will attempt to map the description of structural data onto a set of coordinates that are as abstract as possible, while retaining enough information such that predictions can be used directly to perform a calculation.

**Notation and set-up** Our goal is to make predictions about the likelihood of structures appearing at low temperature and pressure as a function of composition. For this purpose, consider a system composed of $m$ elements denoted $e_1, e_2, \ldots, e_m$. We start by discretizing the continuous composition space $(c_1, c_2, \ldots, c_m)^T \equiv \vec{c} \in \{c_i \in [0,1] s.t. \sum_i c_i = 1\}$ into a finite set of $p$ compositions denoted $(\vec{c}_1, \vec{c}_2, \ldots, \vec{c}_p)$. Each of the $p$ compositions is thus an $m$-component tuple such that $\sum_i c_i = 1$. Obviously, choosing to discretize a continuous space is an approximation with potential complications; a discussion of these topics is provided in the next section. To each element $e_i$ and composition $\vec{c}_i$ we will assign a variable. Let the variable $X_{e_1}$ denote the variable indicating *what* the $1^{st}$ constituent is; e.g. Oxygen, Carbon, Thallium, etc. Likewise the variable $X_{\vec{c}_1}$ denotes the *structure prototype* appearing at composition $\vec{c}_1$ or "nothing known" if no information is available. Note that the variable values are of no particular order and are thus of nominal type (rather than ordinal or numerical). The collection of random variables $X_{e_1}, X_{e_2}, \ldots, X_{e_m}, X_{\vec{c}_1}, X_{\vec{c}_2}, \ldots, X_{\vec{c}_p}$ is

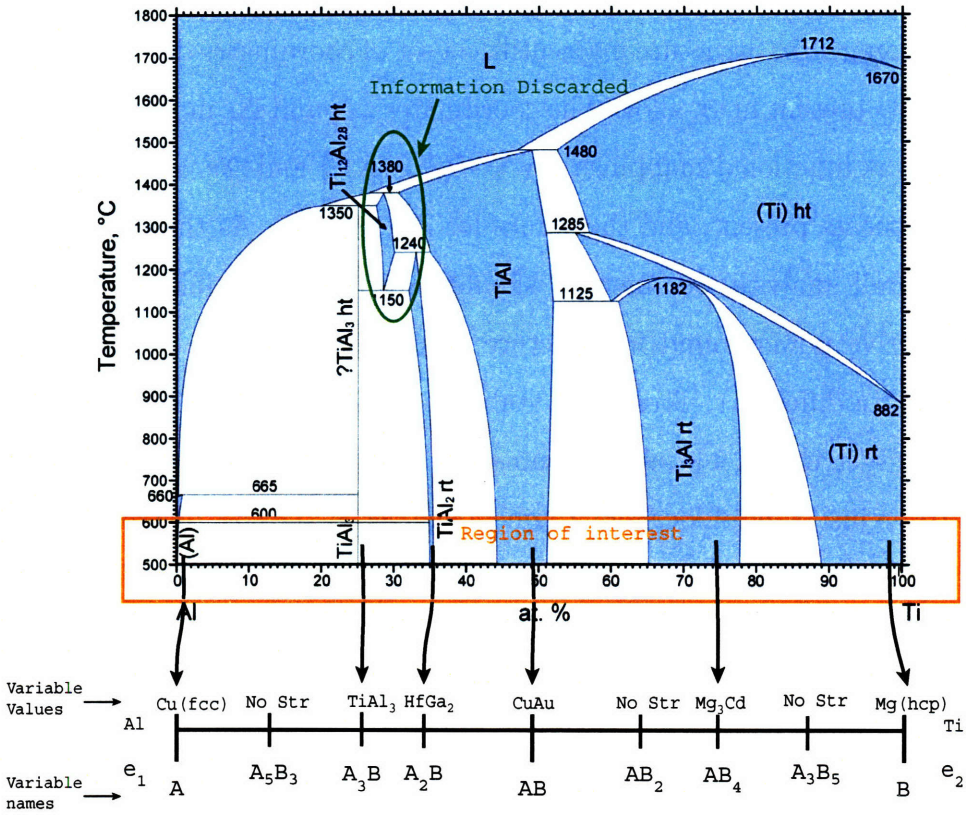


Figure 2-1: Example of how experimental information for the Al-Ti system is mapped onto our set of variables representing what structures are present as a function of composition. (taken from [1])

a $m+p$ tuple denoted by the symbol $\mathbf{X}$. In keeping with common nomenclature, the upper-case $\mathbf{X}$ is used to denote a collection of variables while the lower-case $\mathbf{x}$ will denote a particular *instance* of these random variables. For our purposes, the collection of random variables, $\mathbf{X}$, fully characterizes the low temperature and pressure state of the $m$-component alloy system. A database of information for $N$ alloys, denoted by $\mathcal{D}$, is just a collection of $N$ instances of $\mathbf{X}$ or $\mathcal{D} \equiv \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$. Figure 2-1 demonstrates the relationship between $\mathbf{x}$ and the experimental information available in the Al-Ti phase diagram.

**discussion of variables** In our framework, the set of variables $\mathbf{X}$ fully specifies the information available about an alloy. It is important to contrast this choice of variables with those used in other data mining-like methods such as structure maps discussed

in Section 1.2.2 or methods that try to understand structural stability on the basis of local environments. Structure maps utilize a set of coordinates derived solely from the constituents present in an alloy. The coordinates $\Delta\chi$ and $\Delta r$ in binary alloys can thus be viewed as functional mappings $\Delta\chi = f(X_{e_1}, X_{e_2})$ and $\Delta r = g(X_{e_1}, X_{e_2})$ between the constituents present, and the values for $\Delta\chi$ and $\Delta r$. As these variables are part of our overall analysis, our method should in principle be able to capture much of the predictive power found in structure maps (although *how* we use $X_{e_1}$ and $X_{e_2}$ may be quite different). Structural stability has often been understood through an analysis of a structure's local environments. As mentioned in section 1.2.2 viewing a structure as just a set of local environments comes at the cost of a loss of information. In general, it is simply not possible to uniquely reconstruct a structure on the basis of its local environments alone. By using structure prototypes as our basic set of structural descriptors, we retain all of the details of how environments are connected. Note also that we have not resorted to an oversimplified Hamiltonian or model of mixing. Through this choice of variables, we free our method from having to describe the complicated and subtle physical mechanisms responsible for structural stability; they are simply embedded in the data itself. Nevertheless, in the *representation* we use certain approximations have been made, and it is instructive to comment on them. Perhaps the most important approximation is the use of a discretized composition space. Phases are known to appear at seemingly any composition, so how does one justify the use of a finite number of them ?

**physical arguments** First, our method is intended for use at low temperature where, according to the $3^{rd}$ law of thermodynamics, the composition range over which a phase is stable will shrink to a single point. So each phase, stable at zero Kelvin, will be present at only one stoichiometry, but phases can in principle be present at any stoichiometry (e.g., $A_{99}B$, $A_{98}B_2$, $A_{97}B_3$, ...). Crystalline materials are composed of periodic arrangements of unit cells containing a finite number of atoms. The finite number of atoms in the unit cell, say $N$, implies that only a finite set of rational-valued compositions are possible at zero Kelvin. Empirically it has been found that

the majority of compounds observed in nature occur at compositions where $N$ could be quite small (e.g., AB, $A_2B$, $A_3B$, ABC, $A_2BC$, etc.). For example, in the Pauling File binary alloys database 95% of all compounds present form at just 35 compositions. In the Inorganic Crystal Structure Database (ICSD), 95% of all ternary compounds are distributed over just 200 distinct compositions [1].

**machine learning argument** A stronger argument for using a finite set of compositions can be made from a machine learning perspective. Each alloy maps onto an instance of the set of variables $\mathbf{X}$ and correlations will be extracted from a database of alloys, $\mathcal{D}$. Our only requirement is that we do not destroy correlation through a particular choice of variables (discrete compositions). In the context of our data, this would occur only if two or more compounds present in the same alloy are mapped onto the same composition variable $X_{\vec{c}_i}$. If such a situation occurred quite frequently, we would preclude the possibility of capturing such a correlation. Further discussion of this issue will be taken up in section 2.2, where the specific dataset used in this thesis is described.

## 2.2 Structure correlations

**Binary metallic alloys: the Pauling File dataset** At this point, we have described a procedure for mapping a database of computed or experimental alloy information to a set of variables that will be used to extract correlation. The database that we will use throughout this thesis is derived from The Pauling File, Binaries Edition [2, 87]. The Pauling File contains information on more than 10,000 distinct compounds [2] distributed over 2300 binary alloy systems. This dataset was significantly cleaned by Tibbetts [88] who removed high pressure and temperature entries, duplicate listings, and performed a systematic binning procedure. For initial testing and development, this dataset was restricted to 1335 binary alloys containing

[1] this number is obtained from the raw data, which even includes structures with disorder

[2] a compound in this thesis is synonymous with the concept of a thermodynamic phase: a particular crystal structure appearing at a particular composition.

metallic elements. The compounds appearing in each alloy were mapped onto a set of 29 different composition variables consisting of the following stoichiometries (and their symmetric, $c_i \rightarrow (1 - c_i)$, counterparts): A, $A_9B$, $A_6B$, $A_5B$, $A_4B$, $A_7B_2$, $A_3B$, $A_5B_2$, $A_7B_3$, $A_2B$, $A_5B_3$, $A_3B_2$, $A_4B_3$, $A_5B_4$, and AB. After cleaning, a total of 4256 compounds appearing at intermediate compositions remained. In total, 586 structure prototypes are distributed over 4256 entries.

### 2.2.1 Correlation between structures

In nature, a compound forms at low temperature because the interactions between a material's constituents result in an enthalpy of mixing that is negative $\Delta H_{\text{mix}} < 0$ [3] The structure of the compound that forms depends on many factors, some of which were outlined in Section 1.2.2, making the mapping from atomic properties to a compound's structure extremely difficult, if not impossible, without resorting to quantum mechanics. Nevertheless, it is quite clear that if a compound forms at composition $c_i$, then the most likely structures to form at composition $c_j \neq c_i$ would be related to the *structure* forming at $c_i$. Let's illustrate this idea with an example. If the pure elements A and B both form the fcc structure, and at composition AB the CuAu, or $L1_0$ prototype forms [4], then it seems reasonable to expect that compounds forming at other compositions might also be decorations on an fcc lattice. This line of reasoning represents a heuristic argument utilized for many years to simplify the analysis of intermetallic compounds. Part of our goal is to take this strategy from the realm of mental guidelines, to one that is both quantitative and unbiased.

In our abstraction of experimental alloy data, the variable $X_{c_i}$ [5] indicates the structure prototype, or lack thereof, forming at composition $c_i$. We can analyze the

[3]Strictly speaking, a $\Delta H_{\text{mix}} < 0$ is alone sufficient for compound formation at zero Kelvin. At finite temperature one must also include entropic effects and a compound's stability is no longer determined by simply $\Delta H_{\text{mix}} < 0$ or even $\Delta G_{\text{mix}} < 0$.

[4]the simplest possible ordered arrangement on the fcc lattice with composition AB

[5]the vector symbol on composition has been dropped because we are now only concerned with binary alloys

correlation between structures appearing at different compositions through the ratio

$$g(x_{c_i} = v_l, x_{c_j} = v_m) = \frac{p(x_{c_i} = v_l, x_{c_j} = v_m|\mathcal{D})}{p(x_{c_i} = v_l|\mathcal{D})p(x_{c_j} = v_m|\mathcal{D})} \tag{2.1}$$

Here $p(x_{c_i} = v_l, x_{c_j} = v_m|\mathcal{D})$ represents the probability that the structure prototype $v_l$ forms at composition $c_i$ and prototype $v_m$ forms at composition $c_j$ *in the same alloy* given available data $\mathcal{D}$, whereas $p(x_{c_i} = v_m|\mathcal{D})$ represents the probability that the structure $v_m$ forms at composition $c_i$ given a database of known information. Details of how numerical values are obtained for both are given in Appendix B. Using the definition of a conditional probability we can write $g(x_{c_i}, x_{c_j})$ in two equivalent forms for ease of interpretation.

$$g(x_{c_i}, x_{c_j}) = \frac{p(x_{c_j}|x_{c_i}, \mathcal{D})}{p(x_{c_j}|\mathcal{D})} = \frac{p(x_{c_i}|x_{c_j}, \mathcal{D})}{p(x_{c_i}|\mathcal{D})} \tag{2.2}$$

When $g(x_{c_i} = v_l, x_{c_j} = v_m) > 1$, the structures forming at compositions $c_i$ and $c_j$ are correlated – given that $v_l$ forms at composition $c_i$ it is more likely that $v_m$ will form at composition $c_j$. Likewise, when $0 \leq g(x_{c_i} = v_l, x_{c_j} = v_m) < 1$, these structures are anti-correlated – the presence of one implies that it is less likely to form the other. The correlation ratio $g(x_{c_i}, x_{c_j})$ thus mathematically codifies the concept that structures forming at different compositions can be correlated with one another. However, rather than representing this correlation through a proxy defined by a combination of atomic parameters, it is represented in a *mechanism independent* way. Thus we are not required to explicate a microscopic model detailing *why* the structures are correlated.

Intuitively, highly correlated structures will share common characteristics or conserved units whereas anti-correlated structures will share very little in common. We therefore expect that interactions between a materials constituents will give rise to *conserved* structural motifs between compounds appearing at different compositions. Table 2.1 below gives a few examples of structures that appear correlated in nature as well as a rough summary of the shared features of each. Note that strong correlations are often the result of a dominant bonding mechanism. Thus metal-hydride structures

are highly correlated, as are the so-called "size effect" compounds. Compounds with strong directional bonding, such as $AlB_2$, are also present. Correlations also show up between more traditional intermetallic phases such as $Cu_3Au$ with $Ni_2In$, so this formalism appears to capture the effects of a wide variety of bonding mechanisms. Note that transitive effects are also observed (i.e., if structure A is correlated with structures B and C separately, then B and C are often correlated). For example, in the class of size effect compounds, both $MgCu_2$ and $Fe_3C$ appear often together with the compound $Mn_5C_2$. Incidentally, $MgCu_2$ and $Fe_3C$ are also strongly correlated with one another, having a $g$ factor of 8.5.

One advantage of casting the problem of identifying correlation between crystal structures in a probabilistic fashion, is that we can equally utilize structures that are anti-correlated (i.e., they never appear together). As one might expect, if two structures are stabilized through very different interactions between their constituents, it will be unlikely that they appear together in the same alloy system. These anti-correlations will play a symmetric role in the prediction process, allowing us to rapidly rule out structures as plausible candidates. Table 2.2 gives a few examples of structures that essentially never appear together in nature. As expected, the stability of each prototype in an anti-correlated pair is rationalized through a different mechanism. For example, the $Cu_3Au$ structure is unlikely to appear with $CaF_2$; the former is common to metallic alloys, while the latter appears more frequently in ionic systems.

### 2.2.2 Correlation between compositions: mutual information

Looking at individual correlations between pairs of structure prototypes is useful, but we can take this analysis a bit further in an attempt to answer a slightly more general question. Suppose that you have knowledge of the structure, any structure, appearing at some composition $c_i$. How much does this tell you about the structure appearing at another composition ? Suppose you know what one constituent in an alloy is (but not the others), how should this knowledge change what structures are most likely at

intermediate compositions ? To answer these questions, we would like to know how much information is carried by the outcome of variable $X_{c_i}$ or $X_{e_i}$ with respect to the other variables we have thus defined. This concept is represented mathematically through a quantity called mutual information [91]. Mutual information is a property of two variables and measures an *overall* degree of correlation between the pair. For notational simplicity, let $X_i$ and $X_j$ denote the pair of variables. The mutual information is given by

$$I_{i,j} = \sum_{x_i,x_j} p(x_i, x_j) \ln \left( \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \right) \tag{2.3}$$

Mutual information will fall in the interval $0 \leq I_{i,j} \leq \min(H_i, H_j)$ where $H_i$ is the information entropy of the variable $X_i$ (see Section A.3.2 for further details). When $I_{i,j} = 0$, the variables $X_i$ and $X_j$ are said to be independent or uncorrelated. On the other hand, when $I_{i,j}$ is a maximum, one of the variables is a deterministic function of the other; knowing the outcome of one unambiguously determines the other. In this case $p(x_i|x_j)$ or $p(x_j|x_i)$ becomes a function that is zero or one depending on the combination of $x_i$ and $x_j$. For example, $I_{i,j} = H_i$ will hold when the variable $X_i$ is completely determined through knowledge of $X_j$. Computing values of $I_{i,j}$ for pairs of variables in our analysis will quantify, in a global sense, how much information is carried by knowledge of stable crystal structures. Figure 2-2 shows the mutual information for all pairs of variables in our formalism. Each pixel in the matrix corresponds to the quantity $\frac{I_{i,j}}{\min(H_i,H_j)}$ for various combinations of variables. The strongest correlation in the plot corresponds to the pairs of variables $X_0, X_A$ and $X_1, X_B$ or the mutual information between the "A" or "B" constituents and their ground state structures respectively. Strictly speaking the normalized mutual information between these pairs of variables should be 1; knowing what element is present should uniquely identify its ground state structure at zero temperature and pressure. However, our particular database of information, $\mathcal{D}$, happens to have multiple structural listings available for many elements, resulting in a normalized value that is slightly less than one. The more interesting behavior depicted in Figure 2-2 is the significant degree

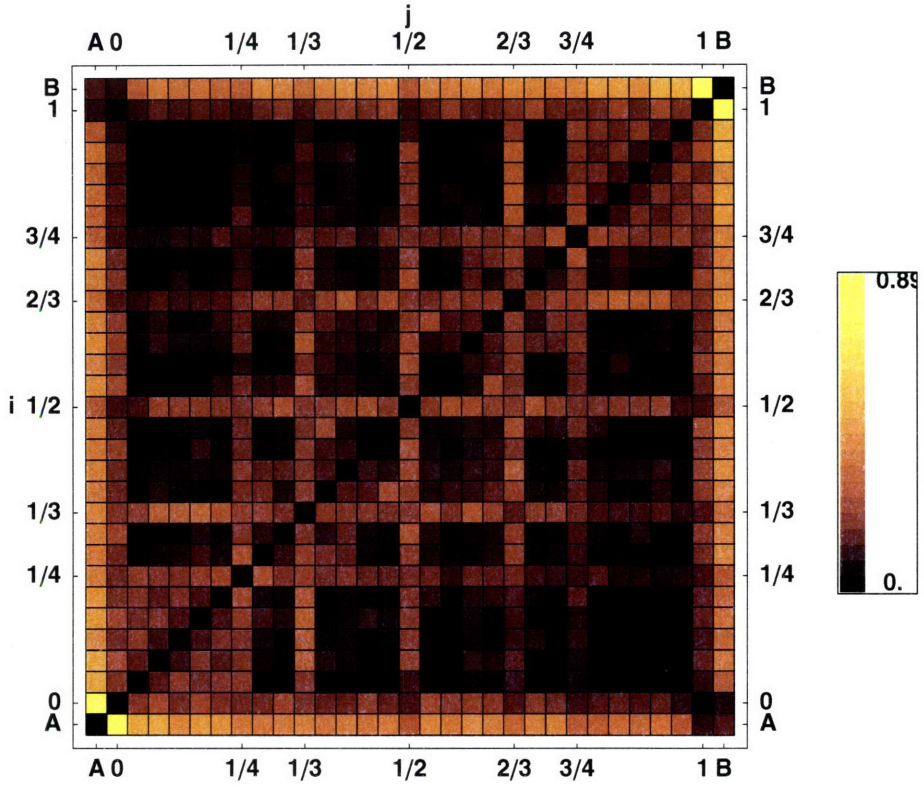


Figure 2-2: Mutual information between pairs of variables using the Pauling File Binaries database. Combinations include pairs of variables where each is a structure forming at intermediate compositions, and combinations involving an element and the crystal structure forming at a particular composition. Lighter colors indicate stronger correlation. (*NOTE-this figure needs to be updated *)

with which pairs of variables are correlated across the composition space. Knowledge of a materials constituents (i.e., knowledge of the variables $X_A$ or $X_B$) clearly has a significant influence on what structures form at intermediate compositions. Some elements, due to the particulars of their electronic structure and how they interact with other elements, tend to stabilize specific crystal structures. For example, binary compounds containing Hydrogen often form the $CaF_2$ structure while those involving Beryllium tend to form the $NaZn_{13}$ structure. These correlations ultimately manifest themselves as the bright rows between the variables $X_A$ or $X_B$ and most of the composition variables. It is also important to note how the bright regions in Figure 2-2 tend to be centered along the diagonal running from bottom-left to top-right. This behav-

ior is to be expected, information about structural stability should have a some degree of locality – the influence of what forms at composition $c_i$ tends to diminish as the focus turns to more "distant" compositions . The strongest composition-composition correlations tend to involve "common" compositions such as $c_i = \frac{1}{3}$ and $c_i = \frac{1}{2}$ (these compositions result in bright rows and columns). Crystal chemistry has long focused on rationalizing the existence of compounds involving a dominant bonding mechanism, such as compounds stabilized by "ionic" versus "covalent" bonds. When a dominant bonding mechanism is present, it is perhaps more likely that a compound will form at a "common" composition such as $c_i = \frac{1}{3}$ or $c_i = \frac{1}{2}$; this information then propagates itself out to other compositions resulting in the observed streaking.

## 2.3 Overall objective: constructing $p(\mathbf{x}|\mathcal{D})$

At this stage of the game, we have (1) developed an abstraction of available crystal structure data, (2) advocated the use of a probabilistic framework for extracting correlations, and (3) shown evidence that such a framework *embodies* many of the rules and guidelines put forth by crystal chemists in the literature. Thus capturing correlation in a probabilistic fashion *implicitly* embeds the physics of structural stability. At no point do we have to connect stability to a microscopic property of a material's constituents, and as such, we avoid many of the pitfalls associated with previously used data mining-like techniques such as structure maps. What remains is to stitch these correlations together into a general technique that can be used for prediction. The central objective of the remainder of this thesis is the construction of a high-dimensional probability distribution $p(x_A, x_B, x_{c_1}, x_{c_2}, \ldots, x_{c_p}|\mathcal{D}) = p(\mathbf{x}|\mathcal{D})$. Before diving into the construction of $p(\mathbf{x}|\mathcal{D})$ let's briefly motivate why such an object would be so useful. Recall our overall strategy discussed in Section 1.3 and shown in Figure 1-2. In the prediction process one will often have some partial degree of knowledge about compounds that are stable in an alloy. As an example, consider the Li-Pt system where the Pt-rich region of the phase diagram is quite well known while the Li-rich region is more uncertain [1]. Missing structural information could arise

for several reasons, such as the difficulty in keeping Li-rich compounds isolated, or because of the large difference between the ability of Li and Pt to scatter radiation (e.g. the atomic scattering factor for Pt is roughly 24 times that of Li for Cu $K_\alpha$ radiation [92]; making matters worse, the measured scattering intensity will go as the square of the structure factor or a ratio of nearly 600 [93]). Thus one is often faced with some partial, and possibly noisy, knowledge about an alloy system and the task is to make predictions about additional phases that could be present. Let the known information, or evidence, be denoted by the symbol **e**. Performing predictions consists of calculating $p(\mathbf{x}|\mathbf{e}, \mathcal{D})$; the probability of $\mathbf{x}$ given the evidence **e**. The known information will, at a minimum, consist of the elements present in an alloy and their ground state structures, but can also include information about phases appearing at intermediate compositions. Based on our discussion of the correlation factors, $g(x_i, x_j)$, and mutual information, it should be clear that highly informed predictions could be made by appropriately using the correlation present in a database of information, $\mathcal{D}$. By constructing $p(\mathbf{x}|\mathcal{D})$ we will be able to utilize the correlations analyzed in Sections 2.2.1 and 2.2.2 to make an informed, logically consistent set of predictions. These predictions will effectively steer detailed quantum mechanical calculations to the most likely set of candidates for a given system.

## 2.4 Graphical models: a framework for building $p(\mathbf{x}|\mathcal{D})$

To construct a probability distribution in the high dimensional space spanned by $\mathbf{x}$ we will make use of a graphical model framework. Graphical models provide a structured approach to building high-dimensional probability distributions where clear dependencies and/or correlations exist between variables of interest. If the variables in our problem, $\mathbf{X}$, were independent we would be able to determine $p(\mathbf{x}|\mathcal{D})$ viz.

$$p(\mathbf{x}|\mathcal{D}) = \prod_i p(x_i|\mathcal{D}) \tag{2.4}$$

However, the evidence presented in Sections 2.2 and 2.2.2 demonstrates that variables in our problem are not independent. In the worst case scenario, one would have to construct a table for each possible value of $p(\mathbf{x}|\mathcal{D})$. However, the number of entries in the table will grow exponentially with the number of variables involved. For example, the Pauling File dataset contains 31 variables each with a sizable domain. Representing $p(\mathbf{x}|\mathcal{D})$ directly would require a table of size $|\Omega_{\mathbf{X}}| - 1 = \prod_i |\Omega_{X_i}| - 1 \sim \mathcal{O}(10^{50})$ which is both impossible to do practically and unnecessary. Rather, the solution to our problem will lie somewhere within the two extremes defined by an independent variable distribution and the full joint probability distribution. Graphical models are the framework of choice for building systematic approximations to high dimensional probability distributions [94]. At their core, graphical models seek to reduce the complexity associated with representing a full joint distribution by *factoring* it into compact terms. Such factorizations can be derived from known independence relationships between variables, but more often they are used to find the best trade-off between representing data and computational complexity. The following section gives a brief overview of graphical models. The purpose here is to highlight only the properties of graphical models needed for our analysis continued in later chapters. Many excellent, and much more thorough, reviews of this material can be found in [95, 96, 94, 97, 98].

Graphical models consist of two parts: (1) a directed or undirected graph over the variables in the problem and (2) an associated probability distribution. The graph is useful for determining and/or representing the qualitative features of the associated probability distribution. Specifically, from the graph one can determine the *independence* properties between random variables that will hold in the associated distribution; e.g. variables $X$ and $Y$ are independent given $Z$ or $(X \perp Y|Z)$.

### 2.4.1 Bayesian Networks

Bayesian Networks consist of a directed acyclic graph (DAG) and an associated probability distribution. The graph, denoted $\mathcal{G}$, consists of a set of nodes and a collection of directed edges between the nodes. Each node in the graph corresponds to a vari-

able, and directed edges explicate dependencies between the variables. To write down the form of a distribution, given $\mathcal{G}$, we need a few definitions. The set of nodes with directed edges feeding into variable $X_i$ are called the *parents* of $X_i$ and the collection of variables associated with these parents is denoted $\mathbf{X}_{Pa_i}$. When no edges feed into variable $X_i$, the set of parents is the empty set, $\emptyset$. Given a set of variables $\mathbf{X} = \{X_1, \ldots, X_m\}$ and a set of parents for each $\{\mathbf{X}_{Pa_1}, \ldots, \mathbf{X}_{Pa_m}\}$, the probability distribution associated with the graph is given by

$$p(\mathbf{x}) = \prod_i p(x_i|\mathbf{x}_{Pa_i}) \tag{2.5}$$

If there are no edges in the graph (i.e., $\mathbf{x}_{Pa_i} = \emptyset$), one is left with the independent variable approximation of Equation 2.4. On the other hand, if the graph is a fully connected DAG, the factorization in Equation 2.5 is equivalent to the product rule (generalization of Equation A.3)

$$p(\mathbf{x}) = \prod_i p(x_i|x_1, \ldots, x_{i-1})$$

which is an identity. Thus, the appearance of edges in the DAG will take us from a set of fully independent variables all the way up to the full joint distribution representing various degrees of correlation along the way. An example of some possible distributions obtained over three variables, and their corresponding graphs are given in Figure 2-3. We can directly read off how a probability distribution will factor given its DAG. Moreover, the structure of a graph determines a set of *independence* relationships that will hold for the associated distribution [94, 95]. For example in Figure 2-3(b) if the outcome of variable $X_1$ is known, the resulting distribution $p(x_2, x_3|x_1)$ factors as $p(x_2, x_3|x_1) = f_2(x_2)f_3(x_3)$ i.e., the variables are conditionally independent given $x_1$ or $(X_2 \perp X_3|X_1)$. However, if the outcome of $x_1$ is not known, the distribution $p(x_1, x_2) = \sum_{x_3} p(x_1, x_2, x_3)$ does not factor and the variables are, in general, correlated. Although we have used equations to demonstrate the independence properties of Figure 2-3(b) one can determine the set of independence state-

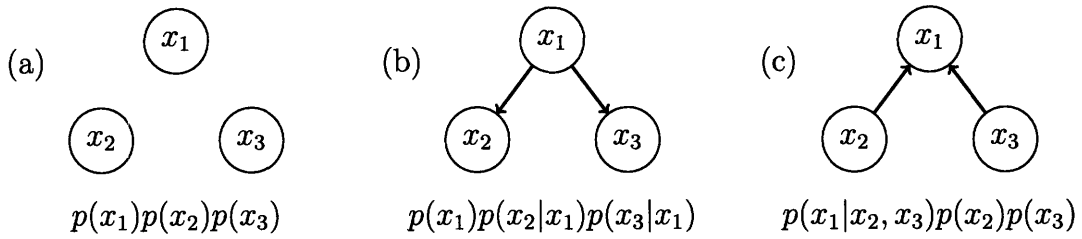


Figure 2-3: Several Bayes Nets over three variables (a) all variables are independent, (b) variables $X_2$ and $X_3$ are conditionally independent given their common parent, $X_1$, and (c) a network where $X_2$ and $X_3$ are marginally independent.

ments that will hold in a DAG through the so-called *d-separation* criterion [94, 95]. Therefore the DAG explicates both the form of the distribution and a set of independence properties that must hold. Due to their use of directed edges Bayesian networks are often most useful in problems where it is possible to identify *causal* relationships between variables. For example, in the field of medical diagnosis a set of variables describing a patient's symptoms will be determined through variables describing a state of disease (i.e., the disease state causes a set of symptoms, not the other way around).

Despite their utility, Bayesian networks have several well-known limitations [95]. For example, given a set of independence statements about a collection of variables, it is not possible in general to uniquely construct a DAG which satisfies them (no solution, a single solution, or multiple solutions may be found). As a consequence, some probability distributions cannot be represented with any Bayesian network whereas others are representable in multiple ways, i.e., multiple graphs each with a different set of directed edges encode the same relationships. Therefore, given *only* a set of data, $\mathcal{D}$, it is only possible to determine the DAG structure up to an equivalence class (the

set of DAGs each making the same independence statements). Interpreting causal relationships from the resulting directed edges between variables therefore must be done with care. A second and related shortcoming of Bayesian networks is the use of directed edges which tend to assign a *direction* of influence. In some settings, the semantics of the problem imply no direction of influence. For example, in our problem the distribution $p(x_{c_i}|x_A, x_B)$ seems reasonable; the appearance of structures is conditioned on their constituent elements. However, the distribution $p(x_{c_i}|x_{c_j})$ is harder to rationalize as we would have to define a "direction" of influence between compositions.

Given a dataset, $\mathcal{D}$, and the functional form of a distribution analytic solutions are available for its parameters [99](see also Appendix B). Moreover, efficient algorithms exist for performing the plethora of *inference* tasks, such as marginalization and prediction. It is even possible to search over the space of factorizations (i.e., the space of graphs $\mathcal{G}$) to select the best model for a given set of data [99]. Summary of the properties of Bayesian networks:

1. Every graph implies a set of independence statements about its variables that *could* be true in the given distribution. Thus a graph indicates the possible independence properties as well as eliminates those that *cannot* be true.

2. Multiple graphs can map onto the same set of independence statements even though their arcs may point in opposite directions. Graphs sharing the same independence statements are said to be equivalent. Without additional information on how data was collected it will be impossible to distinguish between the two.

3. Parameters can be estimated for Bayesian networks very efficiently, in the same way as Appendix B.

### 2.4.2 Markov Networks

Markov networks (or undirected graphical models) [100] are the undirected counterpart of a Bayesian network. A Markov network, like a Bayesian network, is made up

of a graph and an associated probability distribution. Unlike a Bayesian network, the edges connecting nodes in a Markov network are undirected; they represent a generic correlation between the two variables. Discovering the dependence relationships between variables in a Markov network can be obtained simply through the separation properties of the graph [95, 94]. The independence relationships implied by the graph place constraints on the associated distribution, and the *Hammersly-Clifford* theorem [94] provides the connection between independence statements and factorization. Specifically, the theorem states that any associated distribution consistent with the graph *must* factor as

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c) \tag{2.6}$$

where $\mathcal{C}$ is a set of maximal *cliques* [6] present in the graph, $\mathbf{x}_c$ is the subset of variables corresponding to the clique $c \in \mathcal{C}$, and $\psi_c(\mathbf{x}_c)$ is a *potential* function defined over the subset of variables $\mathbf{X}_c$. The quantity, $Z$, is an overall normalization constant sometimes called the partition function of the distribution. The only requirement of a potential function is that it is strictly positive $\psi_c(\mathbf{x}_c) > 0 \; \forall \; \mathbf{x}_c \in \Omega_{\mathbf{x}_c}$.

### relationships between Bayesian and Markov networks

It is important to note some relationships between Markov and Bayesian networks. First, every strictly positive Bayesian network (i.e., $p(\mathbf{x}) > 0$) can be written as a Markov network; the conditional distributions satisfy positivity and the $\psi$'s can be written in terms of the local conditional distributions. There are however Markov networks that cannot be written as a Bayesian network. Consider the Markov network shown in Figure 2-4(a). On the basis of the separation properties of the graph, we can conclude that the variables $X_1$ and $X_3$ are independent given the outcome of $X_2$ and $X_4$ or $(X_1 \perp X_3 | X_2, X_4)$. In addition, we can say that $X_2$ and $X_4$ are independent given the outcome of $X_1$ and $X_3$. These two independence statements are impossible to represent with a Bayesian network over the same variables. Although every Bayesian network *can* be represented as a Markov network, doing so

[6] A clique is a fully connected set of nodes (variables) in the graph and "maximal" means that $c$ is not a sub-set of any other clique in the graph

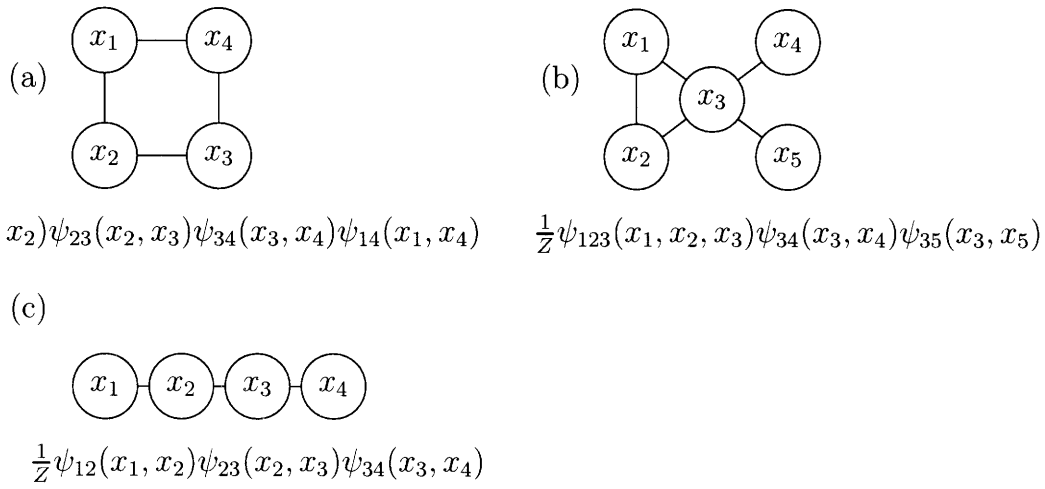


Figure 2-4: Example undirected graphs and their associated probability distributions.

will generally obscure independence statements that can be derived from the directed graph. For example, the $v$-structure shown in Figure 2-3(c) implies that $(X_2 \perp X_3)$, but $(X_2 \perp X_3|X_1)$ is **not** true. These two independence statements cannot be represented with an undirected graphical model using only *two* edges. Rather, the Markov network equivalent of Figure 2-3(c) contains three edges or $p(x_1, x_2, x_3) = \frac{1}{Z}\psi_{123}(x_1, x_2, x_3)$ which obscures the fact that $(X_2 \perp X_3)$ in the Bayes net.

### Parameter estimation

Unlike Bayesian networks, there is often no simple closed-form solution for the parameters of a Markov network (the numerical values returned by the $\psi$'s). The underlying reasons for this difficulty lie in the fact that undirected graphical models do not ascribe a direction of influence between variables, and that parameter estimation necessitates a calculation of the partition function, $Z$, which is intractable to compute in many cases. Thus given a dataset, $\mathcal{D}$, and an undirected graph, it is often considerably

more difficult to obtain appropriate values for the $\psi$'s. An notable exception to this property of Markov networks occurs when the underlying graph is a *tree* [94, 97] in which case the distribution can be written as

$$p(\mathbf{x}) = \frac{1}{Z}\prod_i p(x_i) \prod_{(j,k)\in\mathcal{E}} \frac{p(x_j, x_k)}{p(x_j)p(x_k)} \tag{2.7}$$

where $\mathcal{E}$ is the set of edges appearing in the tree. When the underlying graph is a tree (i.e., it contains no cycles), the parameter estimation process can be carried out in closed form and efficient inference algorithms are available.

### Factor graphs

One drawback of the undirected graphical models discussed thus far is the use of potentials defined over cliques. For example, an undirected graph may become *densely* connected by analyzing just pairwise relationships between variables; i.e., you know that many pairs of variables are correlated, but haven't studied higher order correlations to ascertain their dependencies. For each pairwise correlation, you add an edge to the undirected graph, and before long you have a maximal clique that contains *all* the variables ! The Hammersly-Clifford theorem is useful for making general statements; i.e., statements that will hold for *any* distribution consistent with a given graph. Nevertheless, your distribution of interest could factor in a simpler way, which is certainly allowed. For example, the Hammersly-Clifford theorem indicates that the graph shown in Figure 2-4(b) must factor as $p(\mathbf{x}) = \frac{1}{Z}\psi_{123}(x_1, x_2, x_3)\psi_{34}(x_3, x_4)\psi_{35}(x_3, x_5)$. It may be that your model will only have factors defined over pairs of variables (e.g., one for each edge) in which case the distribution would be written

$$p(\mathbf{x}) = \frac{1}{Z}\psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{23}(x_2, x_3)\psi_{34}(x_3, x_4)\psi_{35}(x_3, x_5)$$

Consider the implication, in terms of storage, of this additional factorization. Suppose for example, each of the 5 variables involved can take on $q$ different values. The

number of parameters required for the Hammersly-Clifford version of $p(x_1, \ldots, x_5)$ is $q^3 + 3q^2$ whereas for the reduced version it is $5q^2$. Hence the reduced factorization will require fewer parameters for any $q > 3$. For cliques containing a large number of variables, this additional factorization would lead to a dramatic reduction in the number of parameters.

Undirected graphical models where the explicit form of the clique potentials has been simplified further, or specified at a finer level, are called *factor graphs.* Every factor graph is an undirected graph, but they are reserved distinction because they are generally simpler.

## 2.5 Outline

The remainder of this thesis focuses on the construction and testing of three different graphical models for $p(\mathbf{x}|\mathcal{D})$. Chapter 3 outlines the construction of a cumulant expansion and Maximum Entropy model for $p(\mathbf{x}|\mathcal{D})$; both undirected graphical models. In Chapter 4 a simple Bayesian network, known as a mixture model will be constructed and analyzed.

| Structure A | Structure B | Correlation Ratio $g(x_{c_i}, x_{c_j})$ | Shared structural elements |
|---|---|---|---|
| $Gd_2Co_7$ @ $\frac{2}{9}$ | $PuNi_3$ @ $\frac{1}{4}$ | 54.0 | Both structures contain the same two local environments for the rare earth ion. The $Gd_2Co_7$ structure arranges these environments in layers with the stacking sequence AABAAB... while in $PuNi_3$ the sequence is ABAB.... |
| $Th_2Ni_{17}$ @ $\frac{2}{19}$ | $PuNi_3$ @ $\frac{1}{4}$ | 30.67 | Both structures share a local environment for the rare earth ion (the "A" environment mentioned above). |
| Ga @ 0 | $AlB_2$ @ $\frac{1}{3}$ | 30.23 | The $AlB_2$ structure contains a Boron layer dominated by $sp^2$ bonding. The $\alpha$-Ga structure is known to contain both covalent and metallic bonds [89]. |
| Ga @ 0 | $PuGa_6$ @ $\frac{1}{7}$ | 46.7 | The 5-fold and 3-fold coordination environments of Ga in $PuGa_6$ are influenced by the covalency character of Ga. |
| $PuGa_6$ @ $\frac{1}{7}$ | $AlB_2$ @ $\frac{1}{3}$ | 93.8 | The covalency bonding character of Ga/B influences the structure. Note these structures are separately correlated with the Ga structure prototype. |
| $HoH_3$ @ $\frac{1}{4}$ | $CaF_2$ @ $\frac{1}{3}$ | 54.7 | Both structure prototypes are common among metal hydrides |
| $MgCu_2$ @ $\frac{1}{3}$ | $Mn_5C_2$ @ $\frac{5}{7}$ | 9.6 | Both structures are so-called "size effect" compounds |
| $MgCu_2$ @ $\frac{1}{3}$ | $CaCu_5$ @ $\frac{1}{6}$ | 6.8 | Both are "size effect" compounds |
| $Fe_3C$ @ $\frac{1}{4}$ | $Mn_5C_2$ @ $\frac{2}{7}$ | 27.2 | Both are "size effect" compounds |
| $Cu_3Au$ @ $\frac{1}{4}$ | $Pu_3Pd_5$ @ $\frac{3}{8}$ | 11.0 | Both are fairly common close-packed intermetallics. $Pu_3Pd_5$ can evidently be derived from the CsCl structure type [90]. One can also view it as an ABAB... stacking of distorted close-packed layers with stoichiometries $A_4B_4$ and $A_6B_2$. |
| $Cu_3Au$ @ $\frac{1}{4}$ | $Ni_2In$ @ $\frac{2}{3}$ | 8.6 | $Ni_2In$ (also called $B8_2$) contains close-packed planes much like the 111 planes of an fcc lattice but with a stacking fault. $Cu_3Au$ is a well-known $A_3B$ ordering in the 111 planes of the fcc lattice. |

Table 2.1: Highly correlated structure prototypes present in the Pauling File database

| Structure A | Structure B | Correlation Ratio $g(x_{c_i}, x_{c_j})$ | Shared structural elements |
|---|---|---|---|
| $MgZn_2$ @ $\frac{1}{3}$ | $MgCu_2$ @ $\frac{2}{3}$ | $\approx 0$ | Both structures are "size effect" compounds. For both to appear in the same alloy system, one would have to place "small" atoms on "large" atom sites. |
| NaCl @ $\frac{1}{2}$ | $MgCu_2$ @ $\frac{2}{3}$ | $\approx 0$ | The ionic bonding present in NaCl is not compatible with a size effect compound |
| NaCl @ $\frac{1}{2}$ | $Cu_3Au$ @ $\frac{3}{4}$ | $\approx 0$ | One is stabilized by ionic interactions, while the other is a simple ordering on a close-packed lattice |
| $Cu_3Au$ @ $\frac{1}{4}$ | $CaF_2$ @ $\frac{1}{3}$ | $\approx 0$ | Bonding characteristics in an alloy will not change significantly over such a small composition range |
| $AlB_2$ @ $\frac{1}{3}$ | CsCl @ $\frac{1}{2}$ | $\approx 0$ | An $sp^2/sp^3$ bonding constituent needed to stabilize $AlB_2$ is incompatible with the CsCl structure |
| W(bcc) @ 0 | $Fe_3C$ @ $\frac{3}{4}$ | $\approx 0$ | The concentration of the "metallic" constituent in $Fe_3C$ should be large, whereas this combination would place the metallic element on the C sites |

Table 2.2: Strongly anti-correlated crystal structures in the Pauling File database

# Chapter 3

# Cumulant expansions for structure prediction

This chapter explores the construction and testing of a probability distribution in the space of all possible binary alloy ground states. Chapters 1 and 2 discussed how such an object would be used, and gave preliminary evidence suggesting that a probabilistic approach to predicting structure would be useful. In this chapter the probability distribution, $p(\mathbf{x}|\mathcal{D})$, is expressed as a truncation to an exact expansion for $p(\mathbf{x}|\mathcal{D})$. The truncated expansion, a *model* for $p(\mathbf{x}|\mathcal{D})$, is then used to make predictions in three specific alloy systems and further tested by predicting all known ground states in binary metallic alloys. The expansion for $p(\mathbf{x}|\mathcal{D})$ is shown to perform remarkably well in spite of a number of known formal difficulties. This chapter finishes with a discussion of these formal issues and how one might correct for them.

## 3.1 Introduction

The idea of approximating a multivariate distribution through a factorization into smaller parts has appeared in many different areas of science. Several general strategies were outlined in Sections 2.4.1 and 2.4.2 in the context of probabilistic graphical models. The model presented in this chapter, a so-called cumulant expansion, is a specialized form of an undirected graphical model. This technique for constructing

$p(\mathbf{x}|\mathcal{D})$ is not in common use in the machine learning or graphical model literature, so this section will begin with a review of *how* the cumulant expansion came about. The goal is to give a bit of background as to the source of our appromation for $p(\mathbf{x}|\mathcal{D})$. We'll start by giving an overview of how Morita [101] used a cumulant expansion to derive Kikuchi's [102] Cluster Variation Method (CVM).

In the field of statistical mechanics one is faced with the problem of computing averages over a probability distribution having the form [32]

$$p(\mathbf{x};\beta) = \frac{1}{Z}\exp\left(-\beta H(\mathbf{x})\right) \tag{3.1}$$

where $H(\mathbf{x})$ is a Hamiltonian defined over the microscopic degrees of freedom of a physical system, $\mathbf{x}$ is a set of variables describing the microstate of the system, Z is a normalization constant, and $\beta = (k_B T)^{-1}$. A well-known result of statistical mechanics [32] is that the thermodynamic state of the system can be determined by performing certain averages over the distribution given by Equation 3.1. To accurately represent the energetics of a system it is often necessary to include terms in the Hamiltonian which couple pairs and multiplets of variables. However, doing so makes the analytic calculation of averages over the equilibrium distribution (Equation 3.1) simply impossible. For example, including pairwise interactions in the Hamiltonian, $\{V(x_i, x_j)\}$, with a general dependence on $(x_i, x_j)$ will couple all the variables in the system. Quantities describing the thermodynamic state of the system, such as the partition function, $Z$, or the entropy, $S$

$$Z(\beta) = \sum_{\mathbf{x}} \exp\left(-\beta H(\mathbf{x})\right)$$

$$S(\beta) = -k_B \sum_{\mathbf{x}} p(\mathbf{x}) \ln\left(p(\mathbf{x})\right) = -k_B \left\langle \ln\left(p(\mathbf{x})\right)\right\rangle$$

require a sum over states to be performed that cannot be generally expressed in any simple analytic form. A method for obtaining approximations to the free energy of the system is to start with the relation between the equilibrium free energy, the average

of $H(\mathbf{x})$ and the entropy $S$

$$F(\beta) = \langle H \rangle - TS \tag{3.2}$$

Starting with Equation 3.2 one develops approximations to the $\langle H \rangle$ and $S$ terms by substituting by substituting in a *trial* form for $p(\mathbf{x};\beta)$ affording easily manipulation. By carefully choosing the trial form for $p(\mathbf{x};\beta)$ it is possible to develop accurate approximations for $\langle H \rangle$ and $S$ terms in Equation 3.2. This strategy was perhaps initiated by Kirkwood [103], and later taken up more generally by Kikuchi [102]. Morita [101, 104] was the first to express Kikuchi's method as a cumulant expansion for the entropy term in Equation 3.2 and shed light on the approximation using the *variational* principle of statistical mechanics.

Morita's cumulant expansion for the entropy is given by

$$S = \sum_{\alpha} \tilde{S}_{\alpha} \tag{3.3}$$

where $\tilde{S}_{\alpha}$ is the entropy cumulant associated with a subset of variables $\mathbf{x}_{\alpha}$ and the sum extends over all possible subsets. The entropy associated with the subset of variables $\mathbf{X}_{\alpha}$, denoted $S_{\alpha}$, is given by

$$S_{\alpha} = -k_B \sum_{\mathbf{x}_{\alpha}} p(\mathbf{x}_{\alpha}) \ln \left( p(\mathbf{x}_{\alpha}) \right)$$

One can solve for the entropy cumulants in terms of the entropies of subsets of variables using a Möbius inversion, as suggested by An [105] and Schlijper [106]

$$\tilde{S}_{\alpha} = \sum_{\beta \subseteq \alpha} (-1)^{n_{\beta} - n_{\alpha}} S_{\beta}$$

where $n_{\beta}$ denotes the number of variables in the subset $\mathbf{x}_{\beta}$. For example, when the subset of variables is $\mathbf{x}_{\alpha} = (x_i, x_j)$, the entropy cumulant is given by

$$\tilde{S}_{i,j} = S_{i,j} - S_i - S_j = -I_{i,j}$$

where $I_{i,j}$ is the mutual information between variables $X_i$ and $X_j$. As discussed in Section 2.2.2 and Appendix A.3.2 the mutual information is zero when the variable $X_i$ is independent of $X_j$ or $(X_i \perp X_j)$. For three variables $\mathbf{x}_\alpha = (x_i, x_j, x_k)$ the entropy cumulant is given by

$$\tilde{S}_{i,j,k} = S_{i,j,k} - S_{i,j} - S_{i,k} - S_{j,k} + S_i + S_j + S_k$$

The entropy cumulant $\tilde{S}_{i,j,k}$ is the negative of the so-called interaction information defined by McGill [107]. Morita developed approximate expressions for the entropy (Equation 3.3) by only including terms involving small subsets of variables. This truncation of Equation 3.3 is equivalent to expressing $p(\mathbf{x};\beta)$ as a factorization over compact functions (functions containing a small number of variables). By only retaining a small number of terms in the factorization, it is possible to arrive at expressions for the $\langle H \rangle$ and $S$ terms in Equation 3.2 involving only the compact functions. Following this simplification, the approximate free energy, denoted $\tilde{F}(\beta)$, is minimized in the low dimensional space of factors rather than attempting to determine $\langle H \rangle$ and $S$ directly (i.e., from a sum over states under the distribution given by Equation 3.1). The variational principle of statistical mechanics [32] guarantees that if the trial form for $p(\mathbf{x})$ is a valid probability distribution [1], the inequality $\tilde{F}(\beta) \geq F(\beta)$ will hold. This technique for obtaining approximate free energies is nowadays referred to as the Cluster Variation Method, or CVM. We have included this discussion of the CVM for two reasons. First, this chapter will start by directly using the expansion for $p(\mathbf{x})$ implied by the CVM; hereafter referred to as the *cumulant expansion* approach. This expansion for $p(\mathbf{x})$ is not often found in the machine learning or graphical model literature, so our discussion has been provided to highlight the origin of this technique. Second, in Section 3.3 we will derive a model for $p(\mathbf{x})$ on the basis of the maximum entropy principle; the CVM can be used in this context as well as a means for parameter estimation.

[1] The cumulant expansion is *not* a valid probability distribution. The consequences of this fact are discussed in Section 3.3.1

**Derivation** Quite generally one can consider a factorization of $p(\mathbf{x})$ in the following form

$$p(\mathbf{x}) = \underbrace{\prod_i g_i(x_i)}_{\text{independent vars}} \underbrace{\prod_{j<k} g_{jk}(x_j, x_k)}_{\text{pair correlations}} \underbrace{\prod_{l<m<n} g_{lmn}(x_l, x_m, x_n) \cdots}_{\text{triplet correlations}} \tag{3.4}$$

The factorization in Equation 3.4 extends over all possible groupings of variables and makes no approximation – it is an identity statement. However, Equation 3.4 suggests that we view a decomposition of $p(\mathbf{x})$ first as a product over independent variable terms with correction terms added to capture correlation present between pairs of variables, triplets, and so on. With this in mind, the cumulants (correction terms) are often *defined* recursively. To see how this works, let's start with a distribution over just one variable. For a single variable Equation 3.4 becomes

$$p(x_i) = g_i(x_i)$$

Now, for a distribution over two variables we write out Equation 3.4 substituting in the single-variable result

$$\begin{aligned} p(x_i, x_j) &= g_i(x_i) g_j(x_j) g_{ij}(x_i, x_j) \\ &= p(x_i) p(x_j) g_{ij}(x_i, x_j) \end{aligned} \tag{3.5}$$

Any distribution over two variables can be written in the form of Equation 3.5. The correction term, $g_{ij}(x_i, x_j)$, is therefore given by $g_{ij}(x_i, x_j) = \frac{p(x_i, x_j)}{p(x_i)p(x_j)}$. Continuing in this fashion, the general form for a cumulant over the variables $\mathbf{X}_\alpha$ is given by

$$g_\alpha(\mathbf{x}_\alpha) = \frac{p(\mathbf{x}_\alpha)}{\prod_{\beta \subset \alpha} g_\beta(\mathbf{x}_\beta)} \tag{3.6}$$

where the product in the denominator extends over all possible subsets of $\mathbf{X}_\alpha$. In Chapter 2 pair cumulants $\{g_{ij}(x_i, x_j)\}$ were analyzed to show how pairs of structures are correlated in nature. When $g_{ij}(x_i, x_j) = 1$ the pair of structures is said to be uncorrelated, because in that case $p(x_i, x_j) = p(x_i)p(x_j)$. The higher-order cumulants,

such as $g_{ijk}(x_i, x_j, x_k)$ can be thought of as a generalization of this concept to larger collections of structures. When $g_{ij}(x_i, x_j) \neq 1$ but $g_{ijk}(x_i, x_j, x_k) = 1$, the structures are pairwise correlated, but not correlated as a triple. The principle approximation used in the CVM is to assume that for subsets of variables beyond a certain size or scope [2], $g_\alpha(\mathbf{x}_\alpha) \approx 1 \; \forall \; \mathbf{x}_\alpha \in \Omega_{\mathbf{X}_\alpha}$; hence there is no need to include these terms in the factorization. For example, using the approximation that any cumulant containing more than two variables will be equal to one, Equation 3.4 becomes

$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{Z} \prod_i p(x_i) \prod_{j<k} g_{jk}(x_j, x_k) \\ &= \frac{1}{Z} \prod_i p(x_i) \prod_{j<k} \frac{p(x_j, x_k)}{p(x_j)p(x_k)} \end{aligned} \tag{3.7}$$

Here we have included an additional constant $Z$ to normalize the distribution. Recall our focus is to determine the probability of a set of ground states $\mathbf{x}$ given available data $\mathcal{D}$, or $p(\mathbf{x}|\mathcal{D})$. To do so we will keep the *form* of Equation 3.7, and just use $p(x_i)$'s and $p(x_j, x_k)$'s *estimated* from the data $\mathcal{D}$ or

$$p(\mathbf{x}|\mathcal{D}) = \frac{1}{Z} \prod_i p(x_i|\mathcal{D}) \prod_{j<k} \frac{p(x_j, x_k|\mathcal{D})}{p(x_j|\mathcal{D})p(x_k|\mathcal{D})} \tag{3.8}$$

where $p(x_i|\mathcal{D})$ and $p(x_j, x_k|\mathcal{D})$ are given according to the procedure presented in Section B.1.2. Equation 3.8 is the model used in this chapter for making predictions, hereafter referred to as just the cumulant expansion.

Before discussing the results obtained with Equation 3.8, it is important to draw the correspondence between Equation 3.8 and the graphical models discussed in Section 2.4. In other words, in the context of graphical models, what is the structure of the graph corresponding to Equation 3.8 ? The graph is not directed, as the probability function does not factor according to Equation 2.5. However, we can view

[2] note that size takes on a different meaning when referring to a *generic* set of variables versus those that, due to their interactions, can be arranged onto a lattice with connecting lines representing interactions.

Equation 3.8 as an undirected graphical model where, as before, each variable corresponds to a node in the graph. Each correction term, $g_{ij}(x_i, x_j)$, in Equation 3.7 adds a direct correlation between the variables $X_i$ and $X_j$. Therefore, for every $g_{ij}(x_i, x_j)$ term appearing in Equation 3.7, an edge will be added to the undirected graph connecting the nodes associated with the variables $X_i$ and $X_j$. If correlation terms are included for *all* pairs of variables, as performed here, the resulting graph is fully connected. Based on this graph connectivity the Hammersly-Clifford theorem (Equation 2.6) makes no statement with regard to the factorization of $p(\mathbf{x})$ because the maximal clique for such a graph corresponds to *all* the nodes. However, the cumulant expansion suggests a much stricter factorization; it is more explicit about how the distribution factors because only pairwise terms are included. Models with such strict factorizations are called factor graphical models and their corresponding graphs, factor graphs [95]. To every pair of variables connected in a factor graph, one associates a non-negative potential function, $\psi_{ij}(x_i, x_j)$. The important concept to note here is that the cumulant expansion (Equation 3.8) explicitly gives the *form* of the potential functions appearing in a factor graph. In particular the cumulant expansion indicates that

$$\psi_i(x_i) = p(x_i|\mathcal{D})$$

and

$$\psi_{ij}(x_i, x_j) = \frac{p(x_i, x_j|\mathcal{D})}{p(x_i|\mathcal{D})p(x_j|\mathcal{D})}$$

Because the potential functions can be directly related to the marginals of the full joint distribution $p(\mathbf{x}|\mathcal{D})$, we call the cumulant expansion a *specialized* form of an undirected graphical model. Parameterizing a factor graph in this way is a bold statement, the consequences of which are discussed later in Section 3.3.

## 3.2 Predictions in binary metallic alloys

Using a filtered [3] version of the Pauling File Binaries Edition for $\mathcal{D}$ we have constructed the expansion of Equation 3.8 and this section outlines our main results. We begin by illustrating predictions for a few specific systems following the outline presented in Section 1.3 and Figure 1-2. As a brief reminder, the overall process consists of the following steps. First, information about an alloy is collected from the database, $\mathcal{D}$, to form the evidence $\mathbf{e}$. Evidence about an alloy consists of, at a minimum, the constituents present and their structure prototypes. In many cases, one will also have knowledge about other compounds present in the system at intermediate compositions – these are all collected into the evidence $\mathbf{e}$. Compositions for which nothing is known are assigned the value "no-compound/2-phase". Next, for the composition of interest, the conditional probability $p(x_i|\mathbf{e},\mathcal{D})$ is calculated for all possible prototypes using the formula

$$p(x_i|\mathbf{e},\mathcal{D}) = \frac{1}{Z(\mathbf{e})} p(x_i|\mathcal{D}) \prod_{j \neq i} \frac{p(x_i, x_j = e_j|\mathcal{D})}{p(x_i|\mathcal{D})p(x_j = e_j|\mathcal{D})} \tag{3.9}$$

where

$$Z(\mathbf{e}) = \sum_{x_i} p(x_i|\mathcal{D}) \prod_{j \neq i} \frac{p(x_i, x_j = e_j|\mathcal{D})}{p(x_i|\mathcal{D})p(x_j = e_j|\mathcal{D})}$$

Assigning the value "no-compound/2-phase" to compositions for which nothing is known is used to simplify the calculation of the conditional probability. If these values were not assigned, calculating the conditional probability would require a sum over all unknown variables; an operation with exponential complexity in the number of unknown variables. The values of $p(x_i|\mathbf{e},\mathcal{D})$ are used to generate an ordered list of likely prototypes (ordered by decreasing conditional probability). Finally, detailed DFT calculations are performed according to the ordered list, and the stability of all phases is assessed. Following the detailed discussions below we'll perform a database-wide set of predictions to establish how our method performs in general.

[3] as described in Section 2.2

### 3.2.1 Specific predictions

In this section, the results of performing structure predictions using the model described by Equation 3.8 followed by detailed calculations are presented for three binary metallic alloys: Ag-Mg, Au-Zr, and Li-Pt. Two of these systems, Ag-Mg and Au-Zr, have been studied previously within DFT by Curtarolo, Morgan, and Ceder [3]. In Curtarolo, et al. DFT calculations were performed on 176 different crystal structures (101 prototypes) in 80 binary metallic alloys. This large scale study gave broad evidence establishing the general agreement between modern DFT techniques and available experimental information. In the Ag-Mg and Au-Zr alloy systems, the crystal structures investigated in Curtarolo et al. [3] were not sufficient to fully ascertain the correspondance between DFT and experimental results (for reasons described below). The results presented here help clarify the comparison between experiment and DFT for these two systems by predicting the structure of compounds for which no comparison was possible in Curtarolo's work.

#### Ag-Mg

The phase diagram of Ag-Mg is known to a reasonable level of accuracy [1, 2, 108, 109], though the structural information for a number of compounds remains unknown. In particular, a compound with composition $AgMg_3$ has been reported by Prokoféf [108] to be stable at low temperature, but the structure has not been refined. Figure 3-1(a) shows the experimental information available for the Ag-Mg system in the Pauling File database. We will use this information as the available evidence, $\mathbf{e}$, to condition our predictions. Note that the Pauling File [2] lists the $ZrAl_3$ prototype ($D0_{23}$) for $Ag_3Mg$ while Massalski [1] lists $Cu_3Au$ ($L1_2$). Detailed DFT calculations [3] and a Long Period Superstructure analysis by Kulik, Takeda, and de Fontaine [109] have confirmed the stable state as $ZrAl_3$, so we will assign the $ZrAl_3$ prototype to the $Ag_3Mg$ composition for making predictions. Figure 3-1(b) shows the most likely candidates for the structure of $AgMg_3$, on the basis the conditional probabilities calculated using Equation 3.9. On the basis of these suggestions, DFT calculations were performed

| Composition | Prototype |
|---|---|
| Ag | Cu (f.c.c.) |
| $Ag_3Mg$ | $ZrAl_3$ ($D0_{23}$) |
| AgMg | CsCl (B2) |
| Mg | Mg (h.c.p) |

(a) Summary of experimental information or evidence, **e**

| Rank | Prototype | $p(x_{\frac{3}{4}}\|\mathbf{e})$ |
|---|---|---|
| 1 | $Cu_{2.82}P$ | 0.063951161 |
| 2 | $BiF_3$ | 0.000000298 |
| 3 | $IrAl_3$ | 0.000000175 |
| 4 | $SrPb_3$ | 0.000000037 |
| ... | ... | ... |

(b) Structure candidates for the compound $AgMg_3$

Figure 3-1: Evidence and predictions in the Ag-Mg system. Experimental information taken from [2].

using the top 10 predicted candidate structures along with 26 other common structures (taken from Reference [3]) within the GGA approximation to DFT [4]. Figure 3-2

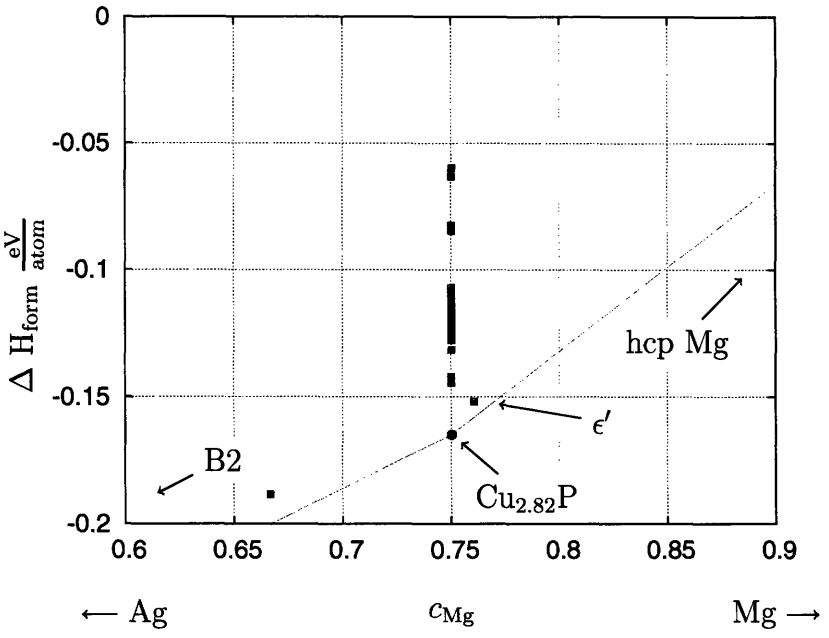


Figure 3-2: Convex hull and DFT calculated formation energies for the Ag-Mg system in the composition range $c_{Mg} \in [0.6, 0.9]$. The tie line on the left hand side connects to AgMg (CsCl prototype) and that on the right to pure Mg (hcp).

[4]Details regarding the convergence parameters used for all DFT calculations performed in this thesis are given in Appendix C)

shows the calculated formation enthalpies (at $p = 0$) of all structures investigated in the present study. Each red square corresponds to the formation enthalpy ($\Delta H_{\text{form}}$) of a structure referenced to the pure end members, in this case face-centered cubic (f.c.c.) Ag and hexagonal close-packed (h.c.p.) Mg. The green line in Figure 3-2, called the convex hull, identifies the thermodynamically stable states of the system as a function of composition. In Figure 3-2, the convex hull extends from the CsCl (or B2) structure at composition AgMg (not shown) to the $Cu_{2.82}P$ structure and from the $Cu_{2.82}P$ structure at $AgMg_3$ to pure Mg (also not shown). The ground states at compositions AgMg, and Mg are well known and both experiment and DFT are in agreement. The $Cu_{2.82}P$ prototype, identified by the cumulant expansion as the most likely compound at the $AgMg_3$ composition, is also the most stable within DFT compound among the 29 structures prototypes calculated at this composition. The $IrAl_3$ prototype, also known as $D0_{18}$, is the next most stable structure relative to the $Cu_{2.82}P$ prototype with an enthalpy 20 $\frac{\text{meV}}{\text{atom}}$ above the hull. The $D0_{19}$ prototype, identified as a potential ground state in Reference [3], lies 22 $\frac{\text{meV}}{\text{atom}}$ above $Cu_{2.82}P$. Also note that an additional phase, the so-called $\epsilon'$ phase, has been experimentally reported at composition $Ag_{17}Mg_{54}$ (off-stoichiometry from the $AgMg_3$ composition). The enthalpy of formation for this phase is also shown in Figure 3-2(just to the right of the $AgMg_3$ composition), but it is unstable with respect to phase separation into a mixture of $AgMg_3$ and pure Mg.

For this prediction, our probabilistic method has been remarkably efficient in suggesting likely ground states. It is important to note that according to practices commonly found in the literature, whereby only common structures are investigated for stability, the $Cu_{2.82}P$ would most likely *not* be tested for stability. This is primarily due the fact that the $Cu_{2.82}P$ structure is both uncommon, appearing a scant nine times in the Pauling File database, and the size of the $Cu_{2.82}P$ unit cell (24 atoms) would discourage its calculation with quantum mechanical methods due to a large computational overhead. In contrast, our method ranks $Cu_{2.82}P$ as a very likely, although non-obvious, candidate on the basis of available experimental correlations.

The ability to predict rare, complicated structure types such as $Cu_{2.82}P$ is an important feature of our method and illustrates that structure correlation can significantly influence the order in which potential ground states are searched for.

### AuZr

The Au-Zr alloy system, like the Ag-Mg system, has been studied within DFT by Curtarolo et al. [3]. A number of ordered compounds have been reported experimentally for this system [1] at the compositions $Au_4Zr$, $Au_3Zr$, $Au_2Zr$, $Au_{10}Zr_7$, $Au_4Zr_5$, $AuZr_2$, and $AuZr_3$. Of these, two compounds $Au_{10}Zr_7$ and $Au_4Zr_5$ lack detailed structural information. In the DFT study [3] of this system no prototypes were calculated at the $Au_{10}Zr_7$ composition. Because no structures were calculated at this composition in Reference [3] the stability of DFT calculated compounds appearing at nearby compositions was inconclusive; it is possible that the $Au_{10}Zr_7$ compound makes compounds appearing at nearby compositions unstable with respect to phase separation. To further clarify this issue, we have performed predictions and DFT calculations for the $Au_{10}Zr_7$ prototype using the available experimental information shown in Figure 3-3(a). A list of suggestions for the structure of $Au_{10}Zr_7$ is generated

| Composition | Prototype |
|---|---|
| Au | Cu (f.c.c.) |
| $Au_4Zr$ | $Au_4Zr$ |
| $Au_3Zr$ | $\beta Cu_3Ti$ ($D0_a$) |
| $Au_2Zr$ | $MoSi_2$ ($C11_b$) |
| $AuZr_2$ | $MoSi_2$ ($C11_b$) |
| $AuZr_3$ | $Cr_3Si$ (A15) |
| Zr | Mg (h.c.p) |

(a) Summary of experimental information or evidence, **e**

| Rank | Prototype | $p(x_{\frac{2}{5}}\|\mathbf{e})$ |
|---|---|---|
| 1 | $Zr_7Ni_{10}$ | 0.00034 |
| 2 | $Os_2Al_3$ | 0.00000 |
| 3 | $Tl_9Pd_{13}$ | 0.00000 |
| 4 | $Cu_5Zn_8$ | 0.00000 |
| 5 | $Ti_2Pd_3$ | 0.00000 |
| ... | ... | ... |

(b) Structure candidates for the compound $Au_{10}Zr_7$

Figure 3-3: Evidence and predictions in the Au-Zr system. Experimental information taken from [2].

with Equation 3.9 based on available evidence, summarized in Figure 3-3(b). In this

study DFT calculations were performed for all structures having a formation energy within $30\frac{\text{meV}}{\text{atom}}$ of the calculated convex hull given in Reference [3]. The convex hull of the system on the basis of these calculations is shown as the green line in Figure 3-4. On the basis of the predictions shown in Figure 3-3(b), we calculated the $Ni_{10}Zr_7$ prototype, suggested as the most likely candidate for this composition. Figure 3-4 shows the calculated convex hull (blue line) *after* the $Ni_{10}Zr_7$ prototype is included into the set of calculated phases. The key observation here is that including the $Ni_{10}Zr_7$ prototype into the set of calculations changes the sequence of DFT-predicted ground states as a function of the Zr composition. This highlights the fact that predicting ground states as a function of composition requires a *thorough* search over the set of possibilities. Constructing the probability distribution, $p(\mathbf{x}|\mathcal{D})$, using a large database of known experimental information enables such a systematic, informed search to be performed.

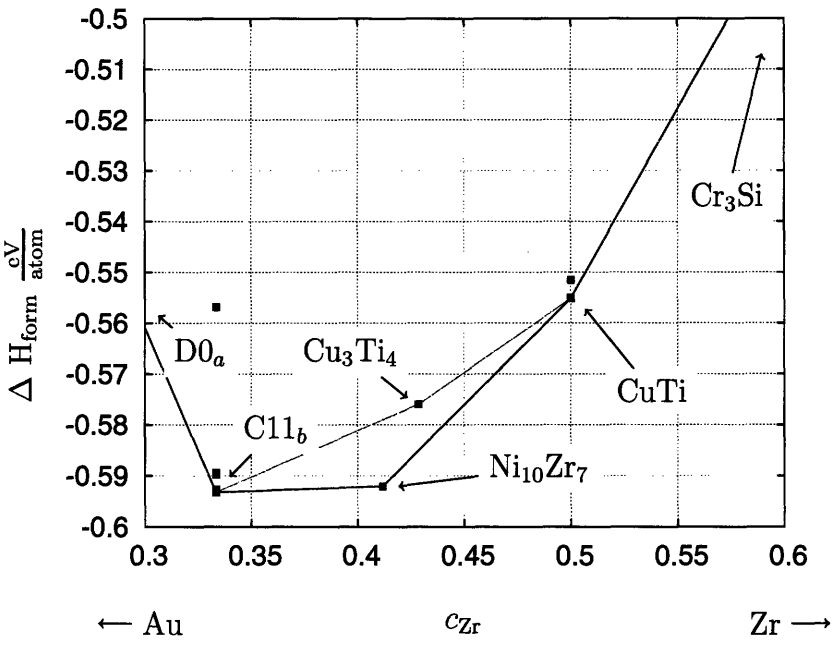


Figure 3-4: Convex hull for the Au-Zr system in composition range $c_{Zr} \in [0.3, 0.6]$. The tie line on the left hand side connects to $Au_3Zr$ ($\beta Cu_3Ti$ prototype) and that on the right to $AuZr_2$ ($MoSi_2$ prototype). The green line corresponds is the calculated convex hull using the results of Reference [3] (without the $Ni_{10}Zr_7$ prototype) while the blue line is the calculated convex hull including the predicted structure for $Au_{10}Zr_7$.

### LiPt

Unlike the Ag-Mg and Au-Zr systems, experimental evidence for the Li-Pt alloy system is both scarce, and occasionally in conflict. The Li-Pt phase diagram in common use [1] is based primarily on the work of Loebich and Raub [110] and supported by only a handful of others [111, 112, 113, 114]. Most of the uncertainty surrounding the compounds of Li and Pt is a result of the experimental difficulties in keeping Li-rich compounds isolated [1]. A good portion of the phase diagram has been called into question, and in particular, little information is available for Li-rich alloys. Experimental structural information is available for the compounds $Li_2Pt$, LiPt, $LiPt_2$, and $LiPt_7$. Two suspected compounds, $Li_5Pt$ and $Li_{15}Pt_4$, have been proposed in the Li-rich portion of the phase diagram on the basis of thermal arrest evidence [111]. To our knowledge, no other DFT calculations have been performed in this alloy system.

Because so little information is known for several compositions in this system, predictions were performed by attempting to maximize the conditional probability $p(\mathbf{x}|\mathbf{e},\mathcal{D})$ over multiple variables at once. In general, the space of possibilities is so large that a fully analytic optimization is not possible. To address this issue a *greedy* search in the space of unknown variables has been performed. Greedy search starts with a random value of $\mathbf{x}$ (i.e., a random assignment to the unknown variables). The function $p(\mathbf{x}|\mathbf{e},\mathcal{D})$ is then optimized, in a sequential manner, over each variable. The variable for which the single largest increase in $p(\mathbf{x}|\mathbf{e},\mathcal{D})$ can be obtained is then fixed, and the procedure repeated. The greedy search algorithm, optimizing one variable at a time, will find a local maximum of $p(\mathbf{x}|\mathbf{e},\mathcal{D})$ in the space of unknown variable assignments. Because only a local maximum is found, the search is repeated several times starting from different random initializations. We have found this procedure particularly efficient for performing predictions over many different variables at once.

Figure 3-5(a) shows the available evidence for the Li-Pt system and Figure 3-5(b), the most likely predictions associated with this system – note that predictions are made at

| Composition | Prototype |
|---|---|
| Li | W (b.c.c.) |
| $Li_2Pt$ | $AlB_2$ (C32) |
| LiPt | LiRh |
| $LiPt_2$ | $Cu_2Mg$ (C15) |
| $LiPt_7$ | $MgPt_7$ |
| Pt | Cu (f.c.c) |

(a) Summary of experimental information or evidence, **e**. The $MgPt_7$ prototype is also known as $CuPt_7$

| Rank | Prototype @ $c_{Pt}$ | $p(\cdot\|\mathbf{e})$ |
|---|---|---|
| 1 | $Cu_3Au@\frac{3}{4}$ | 0.00035289 |
| 2 | $Cu_{15}Si_4@\frac{4}{19}$ | 0.00000680 |
| 3 | $Cu_{2.82}P@\frac{1}{4}$ | 0.00000041 |
| 4 | $LiIr_3@\frac{3}{4}$ | 0.00000034 |
| 5 | $SrPb_3@\frac{3}{4}$ | 0.000000005 |
| 6 | $Sc_{57}Rh_{13}@\frac{13}{70}$ | 0.000000002 |
| ... | ... | ... |

(b) Structure candidates for Li-Pt compounds

Figure 3-5: Evidence and predictions in the Li-Pt system. Experimental information taken from [1, 2].

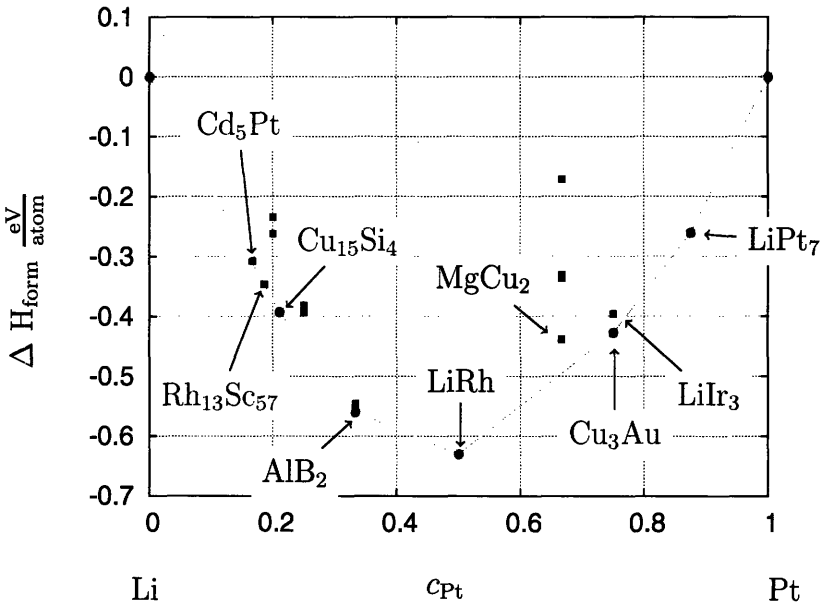


Figure 3-6: Convex hull and DFT calculated formation enthalpies for the Li-Pt system in the full composition range. Ground states are indicated with green diamonds and unstable phases are shown with red dots. Stable phases are labeled by their structure prototype. Blue-faced labels correspond to structures suggested by $p(\mathbf{x}|\mathcal{D})$ while those in black face text correspond to experimentally known phases. The DFT ground state of pure Li is h.c.p. and pure Pt is f.c.c.

multiple compositions. Each prototype listed in Figure 3-5(b) was calculated within DFT along with all experimentally known compounds of Li and Pt[113, 114, 112]. The resulting convex hull and DFT formation enthalpies of all calculated structures are shown in Figure 3-6. Our calculations confirm the stability of the experimentally determined phases $Li_2Pt$, LiPt, and $LiPt_7$. In the Li-rich region of the phase diagram, the suggested structures have led to one ground state prediction at composition $Li_{15}Pt_4$. The formation enthalpies of two additional suggested structures essentially lie on the tie line between pure Li (h.c.p.) and $Li_{15}Pt_4$. In particular, DFT calculations predict the stability of the $Cu_{15}Si_4$ prototype at composition $Li_{15}Pt_4$ and *near* [5] stability of $Li_{57}Pt_{13}$ ( $Rh_{13}Sc_{57}$ prototype) – both compounds were suggested as highly likely candidate ground states by $p(\mathbf{x}|\mathbf{e}, \mathcal{D})$. During the greedy optimization procedure, the cumulant expansion also suggested the $Cu_3Au$, $LiIr_3$, and $SrPb_3$ prototypes as candidates for the compound $LiPt_3$. Our DFT calculations indicate the $Cu_3Au$ prototype is stable for this composition (the structure of $SrPb_3$ is related to $Cu_3Au$ through a tetragonal distortion of the simple cubic unit cell of $Cu_3Au$). The $LiIr_3$ prototype has an enthalpy 32 $\frac{\text{meV}}{\text{atom}}$ above $Cu_3Au$. Note that the experimentally [112] determined structure of $LiPt_2$, the $MgCu_2$ prototype, is found in our calculations to be unstable (by 57 $\frac{\text{meV}}{\text{atom}}$) with respect to phase separation into a mixture of LiPt and $LiPt_3$. It is possible that entropic mechanisms stabilize the $MgCu_2$ structure at finite temperatures over the LiPt and $LiPt_3$ two-phase state, although additional calculations are required to investigate this assertion.

The cumulant expansion has again performed remarkably well in suggesting the ground states of Li-Pt, especially in the Li-rich region of the phase diagram. Calculating only a handful of the most likely suggested candidates is sufficient to identify the structures of several suspected Li-rich compounds. Moreover, note that all structures suggested by the cumulant expansion are either ground states or lie within $35\frac{\text{meV}}{\text{atom}}$ of the calculated convex hull of the system.

[5] this compound has a formation energy that is less than 1 $\frac{\text{meV}}{\text{atom}}$ above the tie line between pure Li and $Li_{15}Pt_4$

### 3.2.2 Database-wide prediction

The successful structure predictions performed in the Ag-Mg, Au-Zr, and Li-Pt systems give some preliminary evidence that our approach to structure prediction is working as desired. However, to get a handle on how well the method performs in general, the cumulant expansion will now be used to predict all compounds in the Pauling File dataset, $\mathcal{D}$. The purpose for this activity is twofold. First, predicting all compounds in the dataset, $\mathcal{D}$, will give an indication of how well the cumulant expansion performs over a wide range of chemistries. Predictions performed in the Ag-Mg, Au-Zr, and Li-Pt systems suggest that the cumulant expansion works well, but how will it perform across a broader class of chemistries ? Second, recall that $p(\mathbf{x}|\mathbf{e})$ gives a ranked list of likely structure prototypes, conditioned on available evidence, $\mathbf{e}$. Once these conditioned probabilities have been determined, structures are calculated with DFT in order of decreasing *likelihood* (i.e., calculate the most probable structures first). The performance of the cumulant expansion is therefore determined by how far down a ranked list one must travel before reaching the true stable structure – i.e., by how effectively the evidence is used to make informed predictions. In practice the position of the true compound is obviously unknown and one is forced to choose a reasonable stopping point. By analyzing the distribution of these positions over all predictions, we will develop an empirical stopping criterion.

Cross-validation [115] is a general technique useful for assessing the predictive power of a model over a dataset. To ascertain both the performance of the cumulant expansion over a wide range of chemistries and obtain a handle on a reasonable stopping criterion for investigating suggested ground states, we have performed cross-validated predictions of all compounds appearing in the dataset, $\mathcal{D}$. The overall outline is given in Algorithm 1. Each compound with a non-unique structure prototype [6] in the database is predicted in the following manner. First, the alloy of interest, denoted $\mathbf{x}_i$,

[6] predicting a unique compound in a cross-validated setting is equivalent to predicting a structure that has never appeared. Unique prototype predictions comprise just 5% of all possible predictions and therefore do not significantly alter the conclusions drawn here.

is removed from the data and the model is re-fit to the new data, $\tilde{\mathcal{D}}$ – this ensures that correlations from the alloy are not used in the prediction process. Next, predictions for the compound of interested will be conditioned using all other information available in the alloy, denoted **e**. The values of $p(x_i|\mathbf{e}, \tilde{\mathcal{D}})$ are then used to generate a sorted list of candidate structures, and the position of the true structure, denoted $l_i$, is recorded from this list. The position of the true compound, $l_i$, is referred to as the *loss* associated with the prediction. In other words, the further down the list one must travel to observe the true structure, the larger the loss. Analyzing how these losses

**Data**: A database of alloys, $\mathcal{D} \equiv \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, and the model $p(\mathbf{x})$
**Result**: A list of losses $\mathcal{L} = \{l_1, l_2, \ldots\}$
**for** *each alloy to be tested* **do**
  `/* Subtract data for alloy` $i$ `from database */`
  form $\tilde{\mathcal{D}} = \mathcal{D} - \mathbf{x}_i$ ;
  fit new model, $p(\mathbf{x}|\tilde{\mathcal{D}})$ ;
  **for** *each compound* $\alpha$ *in alloy* $\mathbf{x}_i$ **do**
    Sort values $x_j \in \Omega_j$ by $p(x_j|\tilde{\mathcal{D}}, \mathbf{e})$ ;
    append position of $\alpha$ to list $\mathcal{L}$ ;
  **end**
**end**

**Algorithm 1**: Procedure for performing cross validated predictions of all compounds in a database, $\mathcal{D}$

are distributed will give an indication of how well the cumulant expansion approach performs on *average*. For example, suppose we want to know how far on average one must travel on the ordered list of prototypes to observe the true compound, i.e., we want to know $\langle l_i \rangle$. We might also be interested in knowing the probability that we have observed the true structure, given that we've descended to a depth $l$ on the ordered list. This information is useful for developing a stopping criterion such as: calculate all structures such that there is at least a $\alpha\%$ chance that the true structure has been calculated. Figure 3.2.2 summarizes the prediction losses for three different approaches to structure prediction; random selection, relative frequency, and the cumulant expansion model presented in this chapter. Note that the cumulant expansion is distinguished through its use of correlation to aid in the prediction process, whereas the other two, random and relative frequency, explicitly ignore correlation.

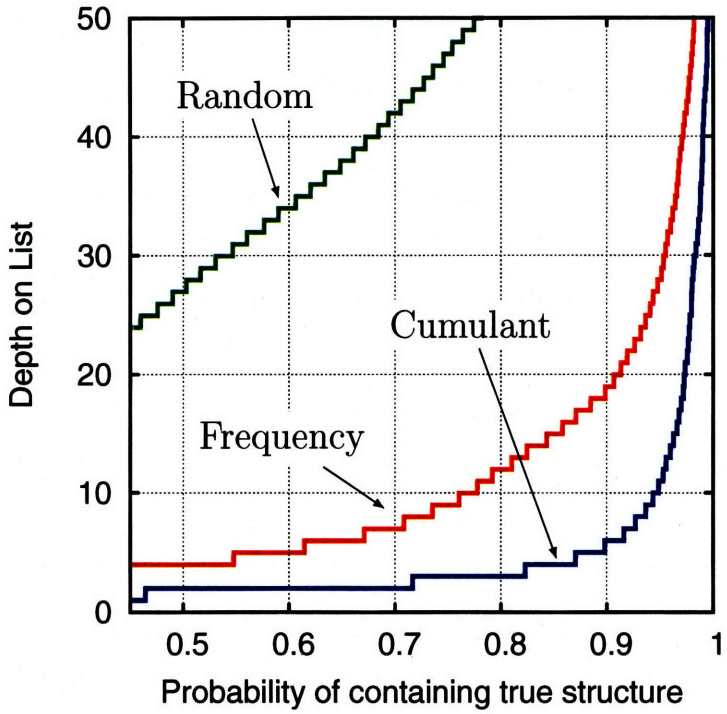


Figure 3-7: Cross validated prediction losses for all compounds in our dataset. Each line indicate the probability that the observed ground state has been seen for a given depth on a sorted list of candidate structures under three different structure suggestion methods: selecting structures at random, by the frequency with which they appear in nature, and according to the cumulant expansion probability model.

Each curve in Figure 3.2.2 shows the depth one must travel on the sorted list to observe the true prototype with a given probability. For example, using the cumulant expansion there is a 90% chance that the true prototype lies within just the first 5 suggested structures. In contrast, picking structures according to the frequency at which they are observed in nature[7] would require a depth of nearly 20 structures to achieve a 90% confidence level. We also note that in more than 40% of the predictions, the cumulant expansion places the true compound at the *top* of the candidate list. Clearly making use of structure-structure correlation has a significant effect on the efficiency with which one would investigate structures for stability. This is especially true in the "high probability" regime of Figure 3.2.2. In this regime the cumulant expansion drastically reduces the number of structures one needs to investigate to have a high confidence in observing the true ground state. These results demonstrate that structure correlations gleaned from historical information can be used to significantly improve the efficiency of searching for stable crystal structures with detailed quantum mechanical calculations. Moreover, because the cross-validation procedure extends over the database as a whole, this conclusion is applicable to a wide range of structures and chemistries. The prediction results shown in Figure 3.2.2 give a general indication of the predictive power of the cumulant expansion, while specific predictions performed in the Ag-Mg, Au-Zr, and Li-Pt alloy systems give a proof-of-principle that the method can efficiently guide quantum mechanical calculations to a relevant, informed structure space.

## 3.3 Refinements and corrections

The cumulant expansion technique presented in this chapter has proven quite useful. However, there are a number of formal problems [116] associated with cumulant expansions that need to be addressed. These formal problems have manifested themselves through a number of related channels. First, note the tables of calculated conditional probabilities, such as the one presented in Figure 3-1(b), are very strongly

[7] stated differently, using the model $p(\mathbf{x}|\mathcal{D}) = \prod_i p(x_i|\mathcal{D})$

peaked around a "most probable" value. In other words, the cumulant expansion appears to be making very strong predictions and the question is whether or not such strong predictions are warranted by the available data. The strongly peaked conditional probabilities suggest that the function $p(\mathbf{x}|\mathcal{D})$ is also strongly peaked in a small region of the space of all possible collections of ground states, denoted $\Omega_{\mathbf{X}}$. A general and consistent measure of how strongly a probability function is peaked or spread out is given by the information entropy of the distribution, denoted $H[p(\mathbf{x}|\mathcal{D})]$ (reviewed in Section A.3). The pairwise cumulant expansion (Equation 3.8), conditioned with the Pauling File binaries edition for $\mathcal{D}$, has an information entropy of just 7.73 bits (obtained through Monte Carlo integration). For comparison, the information entropy of a uniform distribution over the possible ground states is $\ln\left(\prod_i |\Omega_i|\right) \approx 165$ bits, and for the independent variable approximation or "relative frequency" model ($p(\mathbf{x}|\mathcal{D}) = \prod_i p(x_i|\mathcal{D})$) it is $\approx 42$ bits. Information entropy values correspond to the optimal number of binary questions (on average) one would have to ask to determine the outcome of the set of variables $\mathbf{X}$ distributed according to $p(\mathbf{x})$ [96]. The representation of the Pauling File data used in this thesis contains 31 variables in the set $\mathbf{X}$ with $|\Omega_{\mathbf{X}}| \sim \mathcal{O}(10^{50})$ possible outcomes; the information entropy of the cumulant expansion (7.73 bits) implies that a remarkably small number of binary questions are required to determine $\mathbf{x}$ *if the data is truly distributed as the cumulant expansion indicates.* On a final note, the entropy expression for a pairwise cumulant expansion in the CVM formalism (reviewed in Section 3.1) is given by

$$\begin{aligned} S_{CVM} &= \sum_i \tilde{S}_i + \sum_{j<k} \tilde{S}_{j,k} \\ &= \sum_i S_i + \sum_{j<k} (S_{j,k} - S_j - S_k) \\ &= \sum_i H_{X_i} - \sum_{j<k} I_{j,k} \end{aligned} \tag{3.10}$$

Where $H_{X_i}$ is the information entropy for the variable $X_i$ and $I_{j,k}$ is the mutual information (described in Section A.3.2) between the variables $X_j$ and $X_k$. Equation 3.10 corresponds to the Bethe entropy approximation and is negative for the distribution

given by Equation 3.8 using the Pauling File binary alloy data for $\mathcal{D}$. These issues are a result of the same underlying cause; the following sections will attempt to point at the root of the issue, develop a formal fix, and discuss its implementation.

### 3.3.1 The marginalization paradox

Both issues noted above are a result of the fact that the cumulant expansion, as we have used it, *over counts* correlation. We can understand this with a simple example. Suppose one knows the variables $X$, $Y$, and $Z$ are correlated with each other in some manner. If $X$ is correlated to $Y$ and $Y$ to $Z$, then with no other assumptions, $X$ and $Z$ will be correlated through their respective interaction with $Y$. In other words, the mutual correlation of $X$ and $Z$ with $Y$ *induces* correlation between $X$ and $Z$. There is no simple way to explicitly subtract out this induced correlation between $X$ and $Z$ from their true pairwise interaction. To illustrate this a bit further, consider the pairwise cumulant expansion for these three variables–viz.

$$\begin{aligned} p_{CE}(x,y,z) &= \frac{1}{Q}p(x)p(y)p(z)\frac{p(x,y)}{p(x)p(y)}\frac{p(x,z)}{p(x)p(z)}\frac{p(y,z)}{p(y)p(z)} \\ &= \frac{1}{Q}\frac{p(x,y)p(x,z)p(y,z)}{p(x)p(y)p(z)} \end{aligned} \tag{3.11}$$

where $Q$ is an overall normalization constant. Each correlation term in Equation 3.11 includes both "direct" and "induced" correlation rather than just "direct" alone[8]. This fact, by itself, suggests why the cumulant expansion becomes strongly peaked – each correlation term, say for the subset of variables $\mathbf{X}_\alpha$, is tacked onto the expansion as if none of the other variables exists. Now, if this object were a valid probability function, it should *marginalize* properly. By construction, it satisfies the global marginalization constraint $1 = \sum_{x,y,z} p(x,y,z)$, but what about all of the other marginalization constraints ? For example, consider the set of equalities that should

[8]The pairwise cumulant expansion (Equation 3.8) used in this chapter suffers from this correlation overcounting, but in a more severe way due to the number of variables involved.

hold by marginalizing over the variable $Z$.

$$p(x,y) = \sum_z p(x,y,z) \quad \forall\ (x,y) \in \Omega_{X,Y} \tag{3.12}$$

The question is whether or not the cumulant expansion we have used satisfies all marginal consistency constraints required of the distribution $p(x,y,z)$, a subset of which are given by Equation 3.12. The set of all possible marginal consistency constraints applied to a multivariate probability distribution defines what is known as the *marginal polytope* [117, 118]. The marginal polytope describes geometrically, using the marginal probabilities as coordinates, the set of all possible probability distributions defined over a discrete space. Note that the marginal polytope, appearing here in a machine learning context, is conceptually equivalent to the *configurational polytope* used in alloy theory to determine ground states [47, 46, 35, 48]. We can prove that a cumulant expansion will not satisfy these marginal consistency requirements under all possible distributions. For example, substituting Equation 3.11 into Equation 3.12 and simplifying leads to

$$p(x,y) = \frac{1}{Q}\frac{p(x,y)}{p(x)p(y)} \underbrace{\sum_z p(z)p(x|z)p(y|z)}_{\neq p(x)p(y) \text{ in general}} \tag{3.13}$$

The only way for Equation 3.11 to satisfy marginal consistency over $Z$ is for the independence statements $X \perp Z$ and $Y \perp Z$ to hold; these statements are clearly not satisfied in the space of all pair probability functions $p(x,z)$, and $p(y,z)$. Therefore, a cumulant expansion will not satisfy marginalization constraints such as those described by Equation 3.12. It is this marginal inconsistency, in particular, that can lead to unphysical free energies that have occasionally been observed when the Cluster Variation Method has been used in practice [119].

To expand on this a bit further, suppose one is interested in a system containing just three particles with a microstate described by the tuple $x, y, z$. Assuming the

Hamiltonian of the system can be written $H(x,y,z) = h(x,y) + h(x,z) + h(y,z)$ the CVM approximation to the free energy of the system using a pairwise cumulant expansion is given by

$$\begin{aligned}
\tilde{F}(\beta) &= \langle H \rangle - T \sum_{\alpha} \tilde{S}_{\alpha} \\
&= \langle h(x,y) \rangle + \langle h(x,y) \rangle + \langle h(x,y) \rangle - T \left( \tilde{S}_X + \tilde{S}_Y + \tilde{S}_Z + \tilde{S}_{X,Y} + \tilde{S}_{X,Z} + \tilde{S}_{Y,Z} \right) \\
&= \langle h(x,y) \rangle + \langle h(x,y) \rangle + \langle h(x,y) \rangle \\
&\quad - T \left( S_{X,Y} + S_{X,Z} + S_{Y,Z} - S_X - S_Y - S_Z \right)
\end{aligned} \tag{3.14}$$

where, for example

$$\langle h(x,y) \rangle = \sum_{x,y} p(x,y) h(x,y)$$

and

$$S_{X,Y} = -k_B \sum_{x,y} p(x,y) \ln(p(x,y))$$

To obtain an approximation to the free energy of the system one minimizes Equation 3.14 with respect to the probability functions $p(x,y)$, $p(x,z)$, and $p(y,z)$. The key point is that Equation 3.14 is derived *assuming* the cumulant expansion for $p(x,y,z)$ is a valid probability distribution. In particular, to derive the CVM entropy formula it is tacitly assumed that marginalization constraints such as those described by Equation 3.12 are satisified. Because these constraints are not implicitly satisfied by the cumulant expansion, the optimization of Equation 3.14 can result in a set of marginals $\{p(x,y),p(x,z),p(y,z)\}$ which cannot be obtained from *any* probability distribution over $X, Y$ and $Z$. As a consequence, it is no longer possible to claim that $\left[\min_{\{p_{CE}\}} \tilde{F}(\beta)\right]$ is a variational bound to the equilibrium free energy or $\left[\min_{\{p_{CE}\}} \tilde{F}(\beta)\right] > F(\beta)$ – hence, unphysical results are possible.

In the context of this thesis, we refer to the problem described in this section as the *marginalization paradox*. In other words, to construct a cumulant expansion for predicting ground states a set of pair probability functions, estimated from available

data $\mathcal{D}$, is used to form an approximation to $p(\mathbf{x}|\mathcal{D})$. If the resulting function is then used to re-determine the pair probability functions, something very different and inconsistent with the information given is obtained. Using the Pauling File dataset for $\mathcal{D}$ we have validated this assertion on several combinations of highly correlated variables. For example, the mutual information analysis presented in Figure 2-2 indicates that the pairs of variables $(X_A, X_{\frac{1}{4}})$ and $(X_{\frac{1}{4}}, X_{\frac{1}{2}})$ are highly correlated. Checking the marginals of the pairwise cumulant expansion for $p(x_A, x_{\frac{1}{4}}, x_{\frac{1}{2}}|\mathcal{D})$ indicates that roughly 50% of the cumulant expansion's marginals differ from those estimated from the data by more than a factor[9] of 5.

### 3.3.2 A Maximum Entropy approach

Given the cumulant expansion used in this chapter results in a function $p(\mathbf{x}|\mathcal{D})$ that is inconsistent with the data used in its construction, the key question we'd like to answer here is, what *should* the distribution be ? It is possible to solve for the correct distribution by appealing to a principle of maximum entropy – namely, choose the distribution, consistent with known information, that *maximizes* the information entropy over the set of variables $\mathbf{X}$. The motivations behind a maximum entropy approach have been discussed elsewhere[120, 121, 122], but the overall idea is to make $p(\mathbf{x})$ as spread out or non-committal as possible, while remaining consistent with known information. Of all the definitions of "spread out" one could conjure up for probability functions, it turns out that information entropy, $H[p(\mathbf{x})] = -\sum_{\mathbf{x}} p(\mathbf{x}) \ln (p(\mathbf{x}))$, appears to be the most general and theoretically sound [120, 121, 122].

To solve this problem we need to define what "consistent with known information" means mathematically. Note the key ingredient in Equation 3.8 is the set of pairwise probability functions $\{p(x_i, x_j|\mathcal{D})\}$[10]. The pair probability functions encode all the

[9] the marginals of the cumulant expansion are most often *smaller* than those estimated from the data – again, consistent with the idea that the cumulant expansion is erroneously peaked in a small portion of configuration space.

[10] we also use the "point" probability function $p(x_i|\mathcal{D})$, but they are subsumed by the knowledge of $p(x_i, x_j|\mathcal{D})$

correlation used previously to successfully make predictions. The claim here is that these functions, estimated from available data, will serve as our working definition of "known information". Therefore, the solution to our problem can be found by solving for the distribution, $p(\mathbf{x})$, which maximizes the information entropy of $\mathbf{X}$ subject to the condition that it yield pair probabilities consistent with the data. The task now is to translate the above statements into a optimization problem. For this we will make use of the so-called *indicator function* basis or

$$\delta_{i,\alpha} = \begin{cases} 1, \text{ if } x_i = \alpha \\ 0, \text{ otherwise} \end{cases} \tag{3.15}$$

In other words the function $\delta_{i,\alpha}$ is 1 when the outcome of variable $X_i$ is $\alpha$ and zero otherwise. Furthermore, let the vector $\boldsymbol{\phi}_i$ represent the indicator functions associated with the variable $X_i$ or

$$\boldsymbol{\phi}_i^T = \left(\delta_{i,v_1}, \delta_{i,v_2}, \ldots, \delta_{i,v_{\Omega_i}}\right)$$

where $\boldsymbol{\phi}_i^T$ denotes the transpose of $\boldsymbol{\phi}_i$. The vector function $\boldsymbol{\phi}_i(\mathbf{x})$ just indexes the value of $X_i$ in the outcome of all variables, denoted $\mathbf{x}$. For example, suppose that $x_i = v_1$ in $\mathbf{x}$, then

$$\boldsymbol{\phi}_i^T(\mathbf{x}) = (1, 0, \ldots, 0)$$

Because $\boldsymbol{\phi}_i(\mathbf{x})$ only depends on the variable $X_i$ we could use $\boldsymbol{\phi}_i(x_i)$ to denote the same object. In a similar fashion, let $\boldsymbol{\phi}_{i,j}$ denote a vector of indicator functions associated with all combinations of the values of $X_i$ and $X_j$

$$\boldsymbol{\phi}_{i,j}^T = \left(\delta_{i,v_1} * \delta_{j,v_1},\ \delta_{i,v_1} * \delta_{j,v_2}, \ldots,\ \delta_{i,v_{\Omega_i}} * \delta_{j,v_{\Omega_j}}\right)$$

with $\boldsymbol{\phi}_{i,j}(\mathbf{x})$ defined in a similar way. Note that $\delta_{i,v_l} * \delta_{j,v_m} = 1$ when $x_i = v_l$ && $x_j = v_m$ and is zero otherwise. We will let $\langle \boldsymbol{\phi}_i \rangle$ denote the expectation of $\boldsymbol{\phi}_i(\mathbf{x})$ under some given distribution $p(\mathbf{x})$.

$$\langle \boldsymbol{\phi}_i \rangle = \sum_{\mathbf{x}} \boldsymbol{\phi}_i(\mathbf{x}) p(\mathbf{x})$$

Note that $\langle \delta_{i,\alpha} \rangle = p(x_i = \alpha)$.

$$\begin{aligned} \langle \delta_{i,\alpha} \rangle &= \sum_{\mathbf{x}} \delta_{i,\alpha} p(\mathbf{x}) \\ &= \sum_{\mathbf{x}|x_i=\alpha} p(\mathbf{x}) \\ &= p(x_i = \alpha) \end{aligned}$$

Therefore $\langle \boldsymbol{\phi}_i \rangle$ is a vector of the function values for $p(x_i)$ obtained from $p(\mathbf{x})$.

$$\langle \boldsymbol{\phi}_i^T \rangle = (p(x_i = v_1), p(x_i = v_2), \ldots, p(x_i = v_{\Omega_i}))$$

Likewise, $\langle \boldsymbol{\phi}_{i,j} \rangle$ corresponds to a vector of the function values for $p(x_i, x_j)$ also obtained from $p(\mathbf{x})$. Using the functions $\{\boldsymbol{\phi}_i(\mathbf{x})\}$ and $\{\boldsymbol{\phi}_{i,j}(\mathbf{x})\}$ it is possible to determine the point and pair marginals of any valid probability distribution $p(\mathbf{x})$. Our goal is to construct a distribution $p(\mathbf{x})$, defined in the high dimensional space $\mathbf{x} \in \Omega_{\mathbf{X}}$, that is consistent with a given set of low-dimensional point, $\{p(x_i|\mathcal{D})\}$, and pair, $\{p(x_i, x_j|\mathcal{D})\}$, probability functions *estimated* from the given data, $\mathcal{D}$. Let $\boldsymbol{\mu}_i$ denote a vector corresponding to these **estimated** probabilities for variable $X_i$ or

$$\boldsymbol{\mu}_i^T = (p(x_i = v_1|\mathcal{D}), p(x_i = v_2|\mathcal{D}), \ldots, p(x_i = v_{\Omega_i}|\mathcal{D}))$$

Note the distinction; $p(x_i)$ is the marginal of some generic distribution while $p(x_i|\mathcal{D})$ is given as input and based on available data. Our goal is therefore to determine the distribution $p(\mathbf{x})$ such that maximizes $H[p(\mathbf{x})]$ and satisfies the conditions

$$\boldsymbol{\mu}_i = \langle \boldsymbol{\phi}_i \rangle$$

or (element by element)

$$p(x_i = \alpha|\mathcal{D}) = \sum_{\mathbf{x}} \delta_{i,\alpha} p(\mathbf{x}) \quad \forall \quad \alpha \in \Omega_i$$

for each variable. Pair probability constraints are given by the equation

$$\boldsymbol{\mu}_{i,j} = \langle \boldsymbol{\phi}_{i,j} \rangle$$

for each pair of variables, $(X_i, X_j)$. It is worth noting that in general the number of unknown quantities (i.e., the number of probability values we must determine) vastly outnumbers the number of constraints given. Therefore, to solve for the values of $p(\mathbf{x})$ an additional guiding principle (maximum entropy) is required. To maximize information entropy under a set of constraints we begin by forming the Lagrangian

$$\begin{aligned} \mathcal{L}\left(\{p(\mathbf{x})\};\boldsymbol{\lambda}\right) &= H\left[p(\mathbf{x})\right] - \sum_i \boldsymbol{\lambda}_i^T \left(\langle \boldsymbol{\phi}_i \rangle - \boldsymbol{\mu}_i\right) \\ &\quad - \sum_{j<k} \boldsymbol{\lambda}_{j,k}^T \left(\langle \boldsymbol{\phi}_{j,k} \rangle - \boldsymbol{\mu}_{j,k}\right) - \lambda_0 \left(\sum_{\mathbf{x}} p(\boldsymbol{x}) - 1\right) \end{aligned} \tag{3.16}$$

where $\boldsymbol{\lambda}_i$ denotes a vector of Lagrangian multipliers for marginal $p(x_i)$'s constraints, i.e.,

$$\boldsymbol{\lambda}_i^T = \left(\lambda_{i,v_1}, \lambda_{i,v_2}, \ldots, \lambda_{i,v_{\Omega_i}}\right)$$

and $\boldsymbol{\lambda}_{j,k}$ a vector of Lagrangian multipliers for marginal $p(x_j, x_k)$'s constriants. The Lagrangian multiplier $\lambda_0$ ensures proper normalization and the symbol $\boldsymbol{\lambda}$ is used to denote all of the Lagrange multipliers collectively [11]. To solve for $p(\mathbf{x})$, we find the extremal value of Equation 3.16 treating the probability values, the $p(\mathbf{x})$'s, as variables. The first order variation of $\mathcal{L}$ holding the Lagrangian multipliers fixed is given by

$$\begin{aligned} d\mathcal{L}\left(\{p(\mathbf{x})\};\boldsymbol{\lambda}\right) &= \sum_{\mathbf{x}'} \left( -1 - \log(p(\mathbf{x}')) - \sum_i \boldsymbol{\lambda}_i^T \boldsymbol{\phi}_i(\mathbf{x}') \right. \\ &\quad \left. - \sum_{j<k} \boldsymbol{\lambda}_{j,k}^T \boldsymbol{\phi}_{j,k}(\mathbf{x}') - \lambda_0 \right) dp(\mathbf{x}') \end{aligned} \tag{3.17}$$

[11] the single variable terms aren't needed, because $p(x_j) = \sum_{x_k} p(x_j, x_k)$, but they are included for comparison with the pairwise cumulant expansion

Setting each partial derivative to zero yields the solution

$$\begin{aligned} p(\mathbf{x};\boldsymbol{\lambda}) &= \underbrace{e^{-\lambda_0-1}}_{Z^{-1}} \exp\left[-\sum_i \boldsymbol{\lambda}_i^T \boldsymbol{\phi}_i(\mathbf{x}) - \sum_{j<k} \boldsymbol{\lambda}_{j,k}^T \boldsymbol{\phi}_{j,k}(\mathbf{x})\right] \\ &= \frac{1}{Z}\prod_i \psi_i(x_i) \prod_{j<k} \psi_{j,k}(x_j, x_k) \end{aligned} \tag{3.18}$$

where

$$\begin{aligned} \psi_i(x_i) &= \exp\left(-\boldsymbol{\lambda}_i^T \boldsymbol{\phi}_i(x_i)\right) \\ &= \exp\left(-\sum_{v_i} \lambda_{i,v_i} \delta_{i,v_i}\right) \end{aligned}$$

and

$$\begin{aligned} \psi_{j,k}(x_j, x_k) &= \exp\left(-\boldsymbol{\lambda}_{j,k}^T \boldsymbol{\phi}_{j,k}(x_j, x_k)\right) \\ &= \exp\left(-\sum_{v_j}\sum_{v_k} \lambda_{j,k,v_j,v_k} \delta_{j,v_j} \delta_{k,v_k}\right) \end{aligned}$$

Note that the extremum found with this procedure is a unique global maximum of $\mathcal{L}(\{p(\mathbf{x})\};\boldsymbol{\lambda})$ because both the information entropy and the constraint set are convex functions of $p(\mathbf{x})$ [123]. The notation $p(\mathbf{x};\boldsymbol{\lambda})$ is used to indicate the probability distribution parametrized, or indexed, by the full set of Lagrangian multipliers $\boldsymbol{\lambda}$. Note that at this point, we have only solved the problem up to an undetermined set of Lagrangian multipliers $\boldsymbol{\lambda}$.

The principle of maximum entropy has led to the *form* of the correct distribution over $\mathbf{X}$. The form of $p(\mathbf{x};\boldsymbol{\lambda})$ is similar to a cumulant expansion; namely it is a product over functions involving only single variables and pairs of variables. However, unlike Equation 3.8, the parameters of the maximum entropy solution, $\boldsymbol{\lambda}$, are not related in any simple way to the point $(\{p(x_i|\mathcal{D})\})$ and pair $(\{p(x_i,x_j|\mathcal{D})\})$ probability functions estimated from available data. Recall that the pairwise cumulant expansion factors such that $\psi_i(x_i) = p(x_i)$ and $\psi_{i,j}(x_i,x_j) = \frac{p(x_i,x_j)}{p(x_i)p(x_j)}$; the maximum

entropy solution makes no such assertion. Equation 3.18 describes a factor graph or undirected graphical model. As mentioned in Section 2.4.2, fitting undirected graphical models is considerably more difficult than directed graphical models, or Bayesian networks. A general outline of how one would solve for $\boldsymbol{\lambda}$ is shown in Algorithm 2.

**Data**: A set of point and pairwise marginals $\{p(x_i|\mathcal{D})\} \cup \{p(x_i, x_j|\mathcal{D})\}$
**Result**: Parameters, $\boldsymbol{\lambda}$, for the model $p(\mathbf{x}; \boldsymbol{\lambda})$
Initialize $\boldsymbol{\lambda}$;
**while** *not converged* **do**
  Calculate expectation values $\{\langle \boldsymbol{\phi}_i \rangle\}$ and $\{\langle \boldsymbol{\phi}_{i,j} \rangle\}$;
  $\boldsymbol{\lambda} \leftarrow$ parameterUpdate $\left(\boldsymbol{\lambda}, \{\langle \boldsymbol{\phi}_i \rangle\}, \{\langle \boldsymbol{\phi}_{i,j} \rangle\}\right)$;
**end**

**Algorithm 2**: Procedure for determining the parameters, $\boldsymbol{\lambda}$, of the maximum entropy solution of Equation 3.18

The process of determining $\boldsymbol{\lambda}$, or parameter estimation, for the maximum entropy model consists of two parts. First, for a fixed set of parameters we must determine the point and pair marginals of $p(\mathbf{x}; \boldsymbol{\lambda})$ (i.e., the expectation values of the $\boldsymbol{\phi}_i$'s and $\boldsymbol{\phi}_{i,j}$'s). If the marginals of $p(\mathbf{x}; \boldsymbol{\lambda})$ differ from those estimated from the data, the Lagrangian multipliers $\boldsymbol{\lambda}$ must be updated. Therefore following the estimation step, the parameters are updated with the function "parameterUpdate". Parameters can be updated through Iterative Proportional Fitting [124], or a standard Conjugate Gradient minimization of the dual entropy function [125]. Unfortunately, the fitting procedure in Algorithm 2 is made tremendously difficult due to the expectation step. Because the pairwise functions $\psi_{i,j}(x_i, x_j) = \exp\left(-\boldsymbol{\lambda}_{i,j}^T \boldsymbol{\phi}_{i,j}(x_i, x_j)\right)$ couple the variables $X_i$ and $X_j$, determining the marginals of $p(\mathbf{x}; \boldsymbol{\lambda})$ is computationally intractable. As a result, one must resort to a variety of *approximate* methods for determining the marginal probabilities of $p(\mathbf{x}; \boldsymbol{\lambda})$ such as Monte Carlo sampling [126], Belief Propagation [100], its generalization called Generalized Belief Propagation[12] (GBP)[127], or more sophisticated schemes such as "tree re-weighted" belief propagation [128]. Because these schemes can only *approximate* the marginals of $p(\mathbf{x}; \boldsymbol{\lambda})$ one cannot guarantee that Algorithm 2 will even converge to the correct answer. However, some

[12]BP corresponds to minimizing a Bethe free energy to determine the marginals of $p(\mathbf{x}; \boldsymbol{\lambda})$ while GBP is just a new name for using the Cluster Variation Method to accomplish the same task.

empirical evidence by Murphy, Weiss, and Jordan [129] along with the theoretical analysis of Wainwright [130] suggests that approximate techniques are effective practical alternatives.

## 3.4 Summary

In this chapter we have come a long way towards the goal of constructing a probability function $p(\mathbf{x})$ defined in the space of all known ground states of binary alloys. Two different approaches to constructing $p(\mathbf{x})$ were outlined; one based on a truncated cumulant expansion (Equation 3.8) and another based on the principle of maximum entropy (Equation 3.18). Through a number of specific predictions in the Ag-Mg, Au-Zr, and Li-Pt alloy systems, we have shown the practical advantages of using $p(\mathbf{x})$ to *suggest* structures to calculate. Moreover, a large scale cross-validated test of the cumulant expansion demonstrates that it is remarkably efficient in suggesting ground states over a wide range of chemistries. Calculating only a handful of the candidate structures it suggests is sufficient to guarantee the true ground state [13] will be known with a high degree of certainty.

However appealing, the cumulant expansions suffer from a number of known formal problems which were analyzed to arrive at the "most correct" [14] model one could construct using only knowledge of the pairwise probability functions $\{p(x_i, x_j|\mathcal{D})\}$ – the maximum entropy solution in Equation 3.18. Unfortunately, fitting the maximum entropy model is tremendously difficult and one has to question whether the additional complexity in doing so will "pay off" in terms of improved predictive capability. When using the structure prototypes available in the Pauling File binaries edition, solving for the parameters in the maximum entropy model requires minimizing an objective function over several hundred thousand variables ! To make matters worse, this minimization procedure must be carried out when the objective function is known

[13] within the set of known binary alloy ground states
[14] in the sense of information entropy

only *approximately* due to the difficulty in calculating the marginals of Equation 3.18. For these reasons a decision not to pursue the maximum entropy solution further has been made. Rather an effective strategy is to simply restrict the interpretation of the cumulant expansion's results. In particular, the prediction results presented in this chapter suggest using the likelihood *order* assigned to candidate structures based on the cumulant expansion, but the conditional probabilities $p(\mathbf{x}|\mathbf{e}, \mathcal{D})$ are meaningful only up to determining this order. Predicting all non-unique compounds appearing in the Pauling File database provided the information necessary to establishing a stopping criterion for investigating candidate structures over this order. For example, to have a 95% confidence in observing the true ground state among known structure prototypes, it is sufficient to calculate the top 11 suggested candidates with DFT.

# Chapter 4

# Mixture models for structure prediction

Chapter 3 of this thesis discussed the implementation and testing of an undirected graphical model for predicting crystal structure. The results were very encouraging – cross-validated predictions of an entire database of compounds (Section 3.2.2) and detailed predictions in specific alloy systems (Section 3.2.1) demonstrate the utility of using a machine learning method to decide *what* to calculate. Although successful, the cumulant expansion presented in Chapter 3 suffers from a number of formal inconsistencies. The principle of maximum entropy (Section 3.3.2) was used to derive a general solution to constructing a probability distribution although the practical difficulties associated with this approach preclude its use at the moment.

This chapter is devoted to exploring the construction of $p(\mathbf{x})$ starting from a very different probabilistic *model*. Before getting into the details of the model a brief motivation of why it will be useful is given here. First, the cumulant expansion contains a very large number of parameters. In Section 2.2.2 the mutual information between pairs of variables was analyzed and the results obtained provided strong evidence that structures appearing at different compositions correlate with one another. Mathematically, these results suggested that correction terms $\{g_{ij}(x_i, x_j)\}$ were needed in the pairwise cumulant expansion (Equation 3.8) to properly account for this correlation.

Each pairwise correlation term, say between variables $X_i$ and $X_j$, requires the calculation of $|\Omega_{X_i}||\Omega_{X_j}| - 1$ parameters where $|\Omega_{X_i}|$ is the number of elements in the domain of variable $X_i$. Suppose now that one were trying to build a cumulant expansion for multi-component systems. The general procedure described in Section 2.1 starts by discretizing composition space and associating a variable to each discrete composition. For an $n_c$-component system, the number of distinct compositions will scale as $\frac{\alpha^{n_c-1}}{n_c!}$ where $\alpha$ is a characteristic number of divisions along each composition coordinate (e.g., in this thesis $\alpha \approx 30$). A pairwise cumulant expansion will generally require correction terms, $g_{ij}(x_i, x_j)$, for each distinct pair of variables; the number of these terms will scale as $\alpha^{2(n_c-1)}$ and the number of parameters in the model will be proportional to $\alpha^{2(n_c-1)}$. An objective of this chapter is to explore whether or not it is possible to construct a *parameter lean* model for $p(\mathbf{x})$. The idea is that a parameter lean model will scale more favorably[1] as our method is applied to multi-component systems where the number of possible compositions increases dramatically. Ideally, a model will be found that achieves a level of predictive capability similar to the cumulant expansion in Chapter 3, using a much smaller number of parameters.

## 4.1 Introduction

The probabilistic model used in this chapter is known as a mixture model. Mixture models are used in a wide variety of machine learning problems, often in the context of *classification*, i.e., you are given data which can be clustered together into a set of groups or classes. To classify data, we will need to introduce a new variable, $J$, ascribing a label to each data point. Occasionally, data is furnished with labels; in other words $\tilde{\mathcal{D}} \equiv \{(\mathbf{x}, j)_1, (\mathbf{x}, j)_2, \ldots, (\mathbf{x}, j)_N\}$. A *supervised learning* problem consists of determining a mapping between any possible value of $\mathbf{x}$ and the label $j$ using the given data $\tilde{\mathcal{D}}$ as a guide. Once a mapping has been determined it is possible to *predict* a data point's class given $\mathbf{x}$ alone. For example, $\mathbf{x}$ could represent information about a credit card transaction with the labels "fraudulent" and "valid". Using

[1] i.e., the ratio of the number parameters in the model to the number of available data points

a database of transactions that are both fraudulent and valid a mapping between $\mathbf{x}$ and $j$ is constructed. When presented with information about a new transaction, the mapping is used to predict if the transaction is fraudulent. Another example of supervised learning is the task of constructing structure maps. In a structure map, each structure prototype becomes a class label, and $\mathbf{x}$ represents a set of coordinates derived from atomic parameters such as electronegativity difference and valence electron concentration. The available data consists of this set of labeled points and the task is to determine a set of boundaries in the space of $\mathbf{x}$ placing each prototype into its own distinct region of space – ideally the result would look similar to Figure 1-1. More often than not, data is supplied without labels, although it may be both possible and useful to group the data into an underlying set of classes for prediction purposes. Classification problems involving unlabeled data are called *unsupervised learning* problems.

In this chapter, the unlabeled alloy data, $\mathcal{D}$, will be viewed as a collection of an underlying set of groups. It isn't necessary to define what these groups are at the moment, but suffice it to say that because each alloy in $\mathcal{D}$ is identified by a sequence of structure prototypes appearing at intermediate compositions, a group of alloys will be connected by the structure prototypes they share in common. A mixture model will represent this class structure in a probabilistic way by "mixing" the predictions for each class on a probability scale. As before, $\mathbf{x}$ represents an observation of the ground states in an alloy. We will assume that $\mathbf{x}$ can come from a set of $m$ different categories. If the type $j$ for each instance $\mathbf{x}$ is known, we could create the distributions $p(\mathbf{x}|j)$ and $p(j)$. Using the product rule for $p(\mathbf{x}, j)$ (Equation A.3) the distribution $p(\mathbf{x})$ is given by

$$p(\mathbf{x}|\theta) = \sum_{j=1}^{m} p(j)p(\mathbf{x}|j) \tag{4.1}$$

Where $\theta$ is defined as the set of parameters sufficient to specify the distributions $p(j)$ and $\{p(\mathbf{x}|j)\}$. In other words, the probability of $\mathbf{x}$ is taken as a mixture of predictions arising from $m$ different class-specific distributions $\{p(\mathbf{x}|j)\}$. Mixture models of this

type arise in a wide variety of problems where one is trying to explain the behavior of an observed set of variables $\mathbf{X}$ when one or more additional variables cannot be observed. For example, suppose we observe a set of symptoms for a patient, but do not know the underlying disease state generating those symptoms [131]. Ultimately, through an analysis of a large number of patients, patterns of similar symptoms will emerge indicating the same underlying state of disease. In this thesis the variables, $\mathbf{X}$, represent the ground state structures of an alloy. In a database of alloys, $\mathcal{D}$, the patterns that emerge are groups of alloys sharing the same or similar structure prototypes as a function of composition. Conceptually, it may be useful to think of alloy classes as representing different "bonding types". For example, many of the heuristic approaches for rationalizing structure stability outlined in Section 1.2.2 define such classes. Alloys in which the atoms are of very different size would constitute a "size effect" class, whereas those with large electronegativity differences an "ionic bonding" class. Each bonding class will tend to stabilize a limited number of structure prototypes as a function of composition; e.g. a "size-effect" class, if present, would probably contain the $Fe_3C$ and $MgCu_2$ structure prototypes. Loosely speaking, predictions for new alloys are made by correlating whatever partial information is presently available with the structure prototype fingerprints for each bonding class. It is important to note that although "bonding type" may be a useful conceptual tool for understanding the class structure in $\mathcal{D}$, the model (Equation 4.1) does not assume these are actually the classes present. In addition an alloy does not have to belong to just one class; rather the predicted ground states can arise from a 50/50 mixture of "ionic" and "size-effect" ground states (or any other proportion for that matter).

### 4.1.1 The naive Bayes model

One piece of Equation 4.1 that has yet to specified are the class-conditional distributions $\{p(\mathbf{x}|j)\}$. By itself the hidden variable $J$ simply appears to make things more complicated because (1) we now have another variable to deal with and (2) it is never actually observed in the data. Mixture models are used in practice because remarkably *compact* class-conditional distributions can be used to represent seem-

ingly arbitrarily complex behavior [132]. In fact, the very purpose of introducing the hidden variable, $J$, is to considerably simplify the representation of alloy data [133]. The class-conditional distribution we use in this thesis is the following

$$p(\mathbf{x}|j) = \prod_i p(x_i|j)$$

so our model in full form is

$$p(\mathbf{x}|\theta) = \sum_{j=1}^{m} p(j) \prod_i p(x_i|j) \tag{4.2}$$

The form $p(\mathbf{x}, j|\theta) = p(j) \prod_i p(x_i|j)$ results in what is known as a naive Bayes classifier. Models of this type are often adopted in settings where using as few a number of parameters as possible is particularly important. The number of parameters needed to specify the distribution in Equation 4.2 is given by

$$|\theta| = (m-1) + \underbrace{\sum_i (|\Omega_{X_i}| - 1)m}_{n \text{ terms}}$$

where $m$ is the number of mixture components, $n$ the number of variables in the model, and $|\Omega_{X_i}|$ is the number of elements in the domain of variable $X_i$ (i.e., the number of different structure prototypes appearing at the composition corresponding to variable $X_i$). In contrast, the cumulant expansion presented in Chapter 3 is determined by a number of parameters equal to

$$|\theta| = \underbrace{\sum_{j<k} \left(|\Omega_{X_j}||\Omega_{X_k}| - 1\right)}_{\frac{n(n-1)}{2} \text{ terms}}$$

i.e., this is the number of pair probability values that need to be determined to compute Equation 3.8. Therefore, if the number of mixture components, $m$, is small, the number of parameters required for a mixture model will be significantly smaller than the pairwise cumulant expansion used in Chapter 3.

### Naive Bayes independence statements

The mixture model given in Equation 4.2 makes a number of independence statements that are worth mentioning. First, recall the $v$-structure shown in Figure 2-3(b) corresponding to the distribution $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1)$. The discussion in Section 2.4.1 demonstrated that for this distribution the independence statement $X_2 \perp X_3|X_1$ is true (the variables $X_2$ and $X_3$ are independent given $X_1$). However, if the variable $X_1$ is not known, it is no longer possible to claim that $X_2$ and $X_3$ are independent. This is because $p(x_2, x_3)$, obtained by summing $p(x_1, x_2, x_3)$ over $x_1$, results in a function that cannot generally be decomposed into a simple product of functions over each variable, i.e., $\sum_{x_1} p(x_1, x_2, x_3) \neq f_2(x_2)f_3(x_3)$. Because of this marginal dependence between $X_2$ and $X_3$, the two variables are correlated with one another through their mutual dependence on $X_1$. With this in mind, we can better understand the dependence relationships between variables in a mixture model. Figure 4-1 shows the directed acyclic graph corresponding to the distribution for $p(\mathbf{x}, j|\theta)$. The key point is that the Naive Bayes mixture model is just a generalization of the $v$-structure appearing in Figure 2-3(b) to a larger number of variables. Due to the independence properties of the $v$-structure in Bayesian networks, pairs of variables in our model become correlated, all within a rather compact model

$$p(\mathbf{x}|\theta) = \sum_{j=1}^{m} p(j) \prod_i p(x_i|j) \neq \prod_i f_i(x_i)$$

In the context of alloys, if the class variable $J$ is interpreted as "bonding type", the independence statements of the naive Bayes model reads – the structure prototypes appearing as a function of composition are independent, given a bonding type.

## 4.1.2 chemical symmetry

Before moving on to discuss the fitting and testing of a mixture model, a minor subtlety with regard to chemical symmetry is discussed here. For binary alloys our model should be symmetric with respect to the transformation $c \rightarrow (1 - c)$; in other

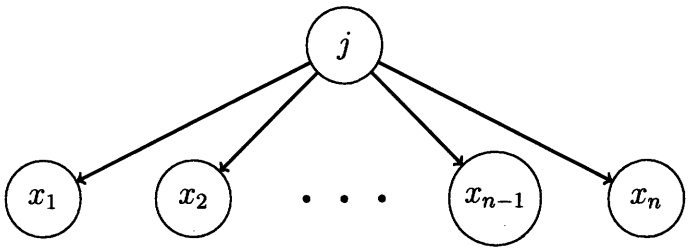


Figure 4-1: Directed acyclic graph corresponding to the naive Bayes model for $p(\mathbf{x}, j|\theta)$.

words if one were to change all compositions in our data from $c \rightarrow (1-c)$ the same model should be obtained. For example, suppose that all alloys in which both the CuAu and $Cu_3Au$ prototypes appear were grouped together into a class. Over the database of alloys, $\mathcal{D}$, this class would consist of alloys in which CuAu appears at $c_B = \frac{1}{2}$ and $Cu_3Au$ appears at $c_B = \frac{1}{4}$ **or** $c_B = \frac{3}{4}$. If no symmetry were used, then two classes would be present; one for the pair CuAu and $Cu_3Au$ appearing at $c_B = \frac{1}{4}$ and *another* for CuAu and $Cu_3Au$ appearing at $c_B = \frac{3}{4}$. An easy way to ensure proper model symmetry is to introduce a permutation variable, $\sigma$, such that Equation 4.2 becomes

$$p(\mathbf{x}|\theta) = \sum_{j=1}^{m} \sum_{\sigma} p(j, \sigma) p(\mathbf{x}|j, \sigma) \tag{4.3}$$

The variable $\sigma$ is used to index a particular permutation or reordering of composition variables. For example, in binary alloys $\sigma \in \{-1, +1\}$ for the AB and BA ordering of elements. Likewise for a $p$-component system, there will be $p!$ different permutations. Note that because $\sigma$ just indexes a permutation of the elements we require that $p(j, \sigma) = p(j, \sigma')\ \forall \sigma, \sigma'$. The class-conditional distributions are now also indexed by

the variables $j$ and $\sigma$

$$p(\mathbf{x}|j,\sigma) = \prod_i p(x_i|j,\sigma) \tag{4.4}$$

For $p$-component alloys only one of the $p!$ class-conditional distributions needs to be independently specified. Each of the other distributions are generated by tracking how variables map onto one another under permutations of the composition variables. For example, in a binary alloy if the variable $X_i$ maps onto $X_j$ under the transformation $c \rightarrow 1 - c$ then we have

$$p(x_i|j,\sigma = +1) = p(x_j|j,\sigma = -1)$$

Therefore, although the sum in Equation 4.3 contains more terms than a "standard" mixture model (Equation 4.2), the number of parameters required to specify a $m$-component "symmetrized" mixture model is no different. In this chapter the primary focus is understanding how to predict crystal structures using a mixture model. We are interested in *what* classes are present in the data and how well they can be used to predict new structures. Therefore, to simplify the notation, we will not make explicit reference to the permutation variable, $\sigma$. However, it should be implicitly understood that chemically symmetrized probability distributions have been used throughout.

## 4.2 Fitting mixture models

The goal of this section is to determine the optimal mixture model given a set of data, $\mathcal{D}$. This process will take place in two stages. First, for a fixed number of components, $m$, the functions $p(j)$ and $\{p(x_i|j)\}$ must be determined i.e., parameter estimation needs to be performed. For a fixed $m$, fitting the functions $p(j)$ and $\{p(x_i|j)\}$ requires care because the variable $J$ is *never* observed. Parameter estimation for mixture models is discussed in Section 4.2.1. Second, provided Equation 4.3 can be reliably parametrized, an optimal number of mixture components, $m_{\text{opt}}$, needs to be determined. Choosing an optimal value for $m$ is a difficult model selection problem discussed in Section 4.2.2.

### 4.2.1 The Expectation Maximization method

For a fixed number of classes, $m$, the model $p(\mathbf{x}|\theta)$ can be fit to data in several different ways. For example, the log-likelihood of the data in terms of $p(\mathbf{x}|\theta)$ is given by

$$ll(\mathcal{D}) = \sum_t \log\left(\sum_j p(j)p(\mathbf{x}_t|j)\right) \tag{4.5}$$

Where $\mathbf{x}_t$ is the $t^{th}$ alloy in the database $\mathcal{D}$. It is possible to optimize Equation 4.5, treating the functions $p(j)$ and $\{p(x_i|j)\}$ as variables, through a standard gradient ascent or conjugate gradient method [133]. However, a method due to Dempster [134], called Expectation Maximization (EM) achieves the same result through an efficient and easy to implement algorithm. The challenge in fitting Equation 4.2 to data is that we do not know the class assignments $j_1, j_2, \ldots, j_N$ *a priori* – i.e., we must determine the labeling of the data points without observing them.

The remainder of this section details the fitting procedure used in the EM procedure. The EM algorithm is briefly outlined as follows. If the complete data, $\tilde{\mathcal{D}} = \{(\mathbf{x}, j)_1, (\mathbf{x}, j)_2, \ldots, (\mathbf{x}, j)_N\}$, were available then the functions $p(j)$ and $\{p(x_i|j)\}$ could be fit directly. Therefore a logical strategy is to use the current setting of the parameters, $\theta$, to estimate class membership (i.e., fill in the missing data). For a given setting of the parameters $\theta$ one calculates the probability that any given alloy $\mathbf{x}$ belongs to the classes $j = 1, 2, \ldots, m$. These conditional probabilities can be thought of as estimating the class labels of the data (for a fixed $\theta$). Using these class labels, the functions $p(j)$ and $\{p(x_i|j)\}$ are then updated, to form a new set of parameters $\theta'$ and the process is repeated until convergence. It can be shown that this process of estimating class membership, followed by forming a new set of parameters increases the log-likelihood of the data at each iteration [134]. A more explicit description of the EM steps is given below.

If the class assignments, $j_1, j_2, \ldots, j_N$ were known, the log-likelihood of the complete

data $\tilde{\mathcal{D}} \equiv \{(\mathbf{x}, j)_1, (\mathbf{x}, j)_2, \ldots, (\mathbf{x}, j)_N\}$ would be given by

$$\begin{aligned} ll(\tilde{\mathcal{D}}) &= \sum_{t=1}^{N} \log\left(p(j_t)\prod_i p(x_i^{(t)}|j_t)\right) \\ &= \sum_{t=1}^{N}\sum_{j=1}^{m} \delta(j|t)\log\left(p(j)\prod_i p(x_i^{(t)}|j)\right) \\ &= \sum_{j=1}^{m}\left(\sum_t \delta(j|t)\right)\log(p(j)) \\ &\quad + \sum_{j=1}^{m}\sum_i\sum_{x_i}\left(\sum_t \delta(j, x_i|t)\right)\log(p(x_i|j)) \end{aligned} \tag{4.6}$$

In Equation 4.6 we let $x_i^{(t)}$ and $j_t$ represent the outcome of variables $X_i$ and $J$ in alloy $t$, $\delta(j|t)$ equals 1 if $j_t = j$ and zero otherwise, and $\delta(j, x_i|t)$ is 1 if $(j_t = j)$ *and* $(x_i = x_i^{(t)})$; zero otherwise. Note we no longer sum over $j$ to obtain the likelihood of the data because the outcome of the variable $J$ is given for each data point. The maximizing solution to Equation 4.6 treating the values of $p(j)$ and $\{p(x_i|j)\}$ as variables (i.e., the parameters) is given by

$$\begin{aligned} p(j) = \hat{\theta}_j &= \frac{(\sum_t \delta(j|t))}{N} = \frac{n(j)}{N} \\ p(x_i|j) = \hat{\theta}_{x_i|j} &= \frac{\sum_t \delta(j, x_i|t)}{\sum_{x_i'}\sum_{t'} \delta(j, x_i'|t')} = \frac{n(x_i, j)}{n(j)} \end{aligned} \tag{4.7}$$

If one were to perform Bayesian estimates of the above quantities, they need to be adjusted according to the development provide in Section B.1.2. Unfortunately, the data is not furnished with labels, and we must discover them automatically. In EM this problem is solved by *guessing* the class assignments using the current parameters, denoted $\theta^{(l)}$. For example, using the parameters $\theta^{(l)}$ the probability that the $t^{th}$ alloy belongs to class $j$ is given by

$$p(j|\mathbf{x}_t, \theta^{(l)}) = \frac{p(j, \mathbf{x}_t|\theta^{(l)})}{p(\mathbf{x}_t|\theta^{(l)})} = \frac{p(j)\prod_i p(x_i^{(t)}|j)}{\sum_k p(k)\prod_i p(x_i^{(t)}|k)}$$

As the conditional probability, $p(j|\mathbf{x}_t, \theta^{(l)})$, represents the probability that alloy $\mathbf{x}_t$ belongs to class $j$, it can be used as a count for $n(j)$ and $\{n(x_i, j)\}$ arising from the $t^{th}$ alloy. To obtain the total counts $n(j)$ and $\{n(x_i, j)\}$, needed for Equation 4.7, one simply adds up the partial counts from each alloy.

$$\begin{aligned} n^{(l)}(j) &= \sum_t p(j|\mathbf{x}_t, \theta^{(l)}) \\ n^{(l)}(j, x_i) &= \sum_t \sum_{x_i} \delta(x_i|t) p(j|\mathbf{x}_t, \theta^{(l)}) \end{aligned} \tag{4.8}$$

where $\delta(x_i|t) = 1$ if $x_i = x_i^{(t)}$ and is zero otherwise. Using the counts determined from the parameters $\theta^{(l)}$, a new set of parameters, $\theta^{(l+1)}$, are determined using Equation 4.7 or its Bayesian counterpart. This two-step process, expectation (Equation 4.8) followed by maximization (Equation 4.7), is iterated until convergence. It can be shown that the EM updates increase the log-likelihood of the data, $ll(\mathcal{D})$, at each iteration [134]. For all results presented in this thesis the initial parameters, $\theta^{(0)}$, are picked from a random distribution. Convergence criteria vary across implementations, but usually one tracks the log-likelihood of the unlabeled data (Equation 4.5) and stops after a tolerance criteria is met. Due to the latent (hidden) nature of the variable $J$, the likelihood surface as a function of the parameters in our model, $\theta$, is littered with local maxima (the likelihood surface is a polynomial containing $m^N$ terms). To circumvent this issue, the EM algorithm is usually run multiple times starting from random initializations. For all results given here, the EM algorithm was restarted at least $200 * m$ times and run for a minimum of $50 * m$ iterations before attempting to evaluate convergence criteria. Figure 4-2 shows the evolution of the log-likelihood function for the Pauling File database as the number of mixing components is changed. For each value of $m$, the EM algorithm was restarted at least $200 * m$ times starting from a random initialization and the model with the best log-likelihood was used for the plot.

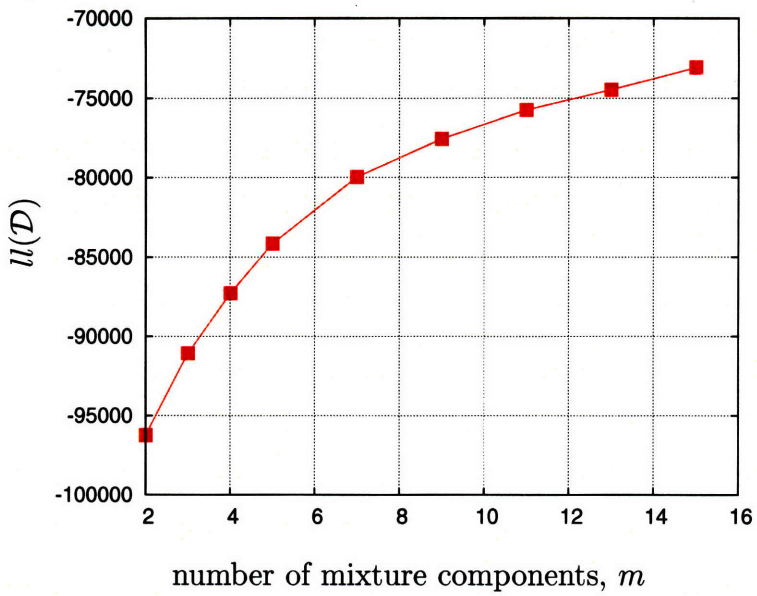


Figure 4-2: Log-likelihood of the Pauling File database of binary alloys as a function of $m$, the number of components in the mixture model. Each point is the maximum likelihood model obtained with $> 200 * m$ restarts of the EM algorithm applied to convergence.

### 4.2.2 Choosing the number of classes

Using the EM algorithm described in Section 4.2.1 it is now possible to obtain the parameters needed for Equation 4.2. To complete the fitting procedure for the mixture model, a sensible number of classes, $m$, needs to be determined. Ideally we will have a criterion to indicate what the "best" number of mixture components is. Unfortunately, the log-likelihood function is of little help here; $ll(\mathcal{D})$ is a non-decreasing function with increasing $m$ [135]. The reason for this behavior is due to the *nesting* property of mixture models. Let $\mathcal{C}_m$ represent the set of all allowed probability functions in the form of Equation 4.1 with $m$ mixture components and $\mathcal{C}_{m+1}$ defined in a similar way. The fact that $\mathcal{C}_m \subseteq \mathcal{C}_{m+1}$, i.e., the model class $\mathcal{C}_m$ is nested within $\mathcal{C}_{m+1}$, leads to the observed behavior in the likelihood as a function of $m$ [135].

Choosing the optimal number of classes is a rather difficult *model selection* problem. The number of classes determines the total number of parameters needed to

specify the distribution, denoted $|\theta|$, which is a surrogate measure of the complexity of the model. Ultimately, we must find an appropriate trade-off between reducing model complexity (favoring a small number of classes) and increasing the likelihood of the data (favoring a large number of classes). A number of techniques have been developed for performing this procedure, each of which tends to fall into two different camps.

**score criteria** One approach for finding the optimal $m$ is to derive an analytic expression, or score, which attempts to balance the two effects of complexity and description of data on the same scale. Therefore, scores typically combine a complexity penalty with the log-likelihood of the data, an indicator of how well the model describes the data [135, 136]. Scoring functions are popular because they can be calculated quickly from a single converged model and typically take the form

$$\text{score}(m, \mathcal{D}, p(\mathbf{x}|\theta)) = \text{complexity}(m, p(\mathbf{x}|\theta)) - ll(\mathcal{D}) \tag{4.9}$$

Where $\text{complexity}(m, p(\mathbf{x}|\theta))$ is a term which is larger for more complex models and $ll(\mathcal{D})$ is the log-likelihood of the data. Scores commonly used in practice are the Bayesian Information Criterion (BIC), Minumum Description Length (MDL), and Minimum Message Length (MML); each using a slightly different form for the $\text{complexity}(m, p(\mathbf{x}|\theta))$ term appearing in Equation 4.9 (see reference [135] for more detail). Complexity functions are often derived from asymptotic behavior in the "infinite data" limit [137, 138, 135, 136]. Thus scores are easy to calculate, but the conditions under which the score becomes valid are often not met in practice. The score used in this chapter is due to Figueiredo and Jain [135] and is given by

$$\begin{aligned} \text{score}(m, \mathcal{D}, p(\mathbf{x}|\theta)) &= \frac{c}{2}\sum_{j=1}^{m}\log\left(\frac{Np(j)}{12}\right) + \frac{m}{2}\log(\frac{N}{12}) \\ &\quad + \frac{m(c+1)}{2} - ll(\mathcal{D}) \end{aligned} \tag{4.10}$$

Where $N$ is the number of data points in $\mathcal{D}$, $m$ the number of mixture components, and $c = m\sum_i (|\Omega_i| - 1)$. Equation 4.10 was picked from the myriad of score options based on the empirical success demonstrated in reference [135]. The score given by Equation 4.10 is rooted in communication theory and is used here primarily for comparison to the re-sampling techniques discussed below. Loosely speaking communication theory treats the model selection problem in the context of communicating the observed data $\mathcal{D}$. By forming the model $p(\mathbf{x}|\theta)$ which describes how data is distributed, it is possible to devise an encoding of the data which is *efficient* (short message). To communicate $\mathcal{D}$ over a channel, one must transmit the data with a message length proportional to $ll(\mathcal{D})$, as well as the model $p(\mathbf{x}|\theta)$ used for encoding. More complex models require longer messages to transmit the encoding scheme leading to the complexity term in Equation 4.10. The key point is that the complexity terms in Equation 4.10, $\frac{m}{2}\log(\frac{N}{12})$ and $\frac{m(c+1)}{2}$, will increase with increasing $m$ to offset the decrease in $-ll(\mathcal{D})$. Note that the complexity terms are linear in $m$ whereas the log-likelihood of the data $ll(\mathcal{D})$ is sub-linear (see Figure 4-2). Therefore, an optimal setting of $m$ can be obtained by *minimizing* Equation 4.10 with respect to $m$.

**re-sampling** An alternative to score-based criteria, called re-sampling methods [135, 115], take the approach of separating the data into two pieces, a *training* set and a *test* set. The model is fit to training data, and its predictive ability evaluated on the test data. In principle as the model complexity surpasses what is warranted by the data its performance on the test data will degrade. The so-called leave one out cross-validation (LOOCV) used in Section 3.2.2 falls into this category of re-sampling methods. Techniques such as LOOCV are useful because they measure the predictive power of a model *directly* on the data available rather than relying on certain limit criteria to be satisfied (as in the score approach). Re-sampling methods are widely used, and provide as a by-product, a quantitative measure of the predictive power of a given model. Therefore, we will adopt a re-sampling technique to determine the optimal number of mixture components.

### 4.2.3 Predictions

Thus far we have (1) given an overview the mixture model, (2) described the independence relationships implied by such a model, and (3) discussed how one would fit Equation 4.1 to available data $\mathcal{D}$. In this section we will present cross-validated prediction results for all non-unique compounds appearing in our dataset $\mathcal{D}$. Our goals are the following:

1. cross-validated predictions will be made for models with a varying numbers of mixture components $m = m_{min}, \ldots, m_{max}$ allowing for the optimal number of mixture components to be determined, $m_{opt}$

2. as a by-product of the prediction steps, we will obtain a database-wide picture of how well a mixture model suggests compounds

Essentially this section performs the same procedure described in Section 3.2.2 with a few minor modifications to Algorithm 1. Recall that to perform a cross-validated prediction, each alloy is first removed from the dataset $\mathcal{D}$, the probability function is then re-fit, and the result is used for prediction. However, re-fitting a mixture model with EM after an alloy is removed from available data (for every alloy) is computationally prohibitive. Each re-fitting step requires many restarts of the EM procedure, resulting in a large computational overhead. So rather than performing LOOCV, the dataset, $\mathcal{D}$, is first split into 20 groups of alloys chosen at random and perform cross-validated tests on each group is performed. Predictions on each group are obtained by first removing the group from the dataset, refitting the mixture model with EM [2], and predicting each compound appearing the random group of alloys. As in Section 3.2.2 we track the performance of our method by monitoring the position of the true compound on a list of predictions – defining a loss $l_i$ for each predicted compound. Figure 4-3 shows the average loss (list position) over all compound predictions in our binary metallic alloy data as a function of the number of mixture components $m$. As expected, increasing the number of mixture components $m$ reduces the average

[2] as before, at least $200 * m$ restarts of the EM algorithm were performed to obtain converged results

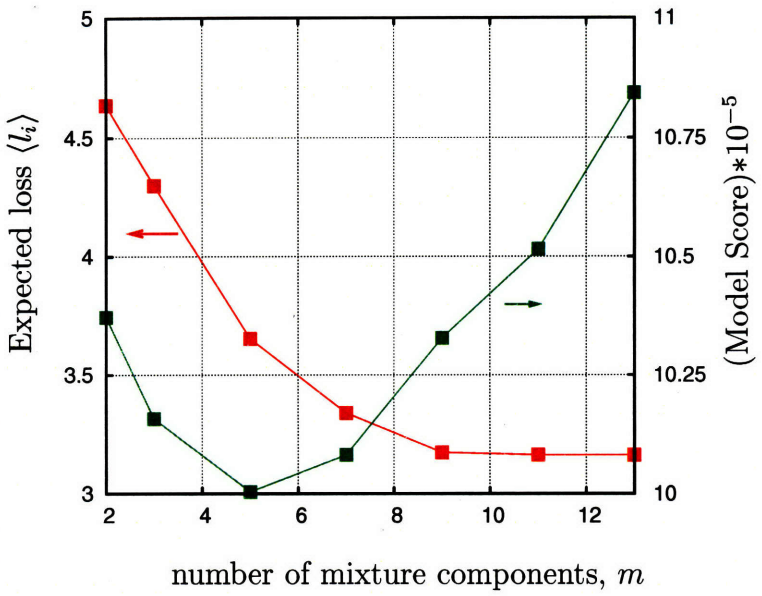


Figure 4-3: Expected losses for a 5% holdout cross-validation test as a function of the number of mixture components $m$. For each $m$, the model score is calculated from Equation 4.10 using a model with the largest $ll(\mathcal{D})$ out of $\approx 200m$ EM restarts.

loss (better predictive capability), but essentially saturates near $m \approx 9$. Figure 4-3 appears to indicate an optimal number of mixture components $m_{opt} = 9, 10$, or 11. Note that a model score (green line in Figure 4-4), given by Equation 4.10, suggests a more conservative setting of $m$. As mentioned previously, model scores are derived analytically in the limit of large $N$, the number of data points. It should therefore be expected that in a data-limited setting (small $N$), the model score will tend to place too much emphasis on the model complexity term in Equation 4.9. Figure 4-4 shows the list length required to contain the true compound with a given probability for three different methods: (1) an ordering based on the prototype's frequency of occurrence in nature (red curve), (2) the cumulant expansion discussed in Chapter 3 (green curve), and (3) a mixture model with $m = 9$ mixing components (blue curve). Each prediction consists of generating a list of candidate prototypes ranked by the conditional probability $p(x_i|\mathbf{e}, \theta)$. On the basis of Figure 4-4 we can conclude that the mixture model has a predictive capability that is approaching the cumulant expansion discussed in Chapter 3. However, as mentioned in Section 4.1.1, the naive

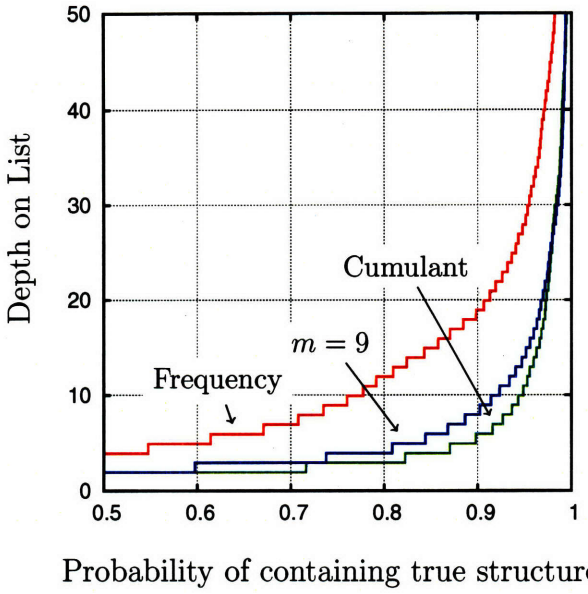


Figure 4-4: List length required to contain the true compound for a given probability. Three different approaches to structure prediction are shown (red curve) picking structures according to their frequency of appearance in nature, (blue curve) a mixture model with $m = 9$ components, and (green curve) the cumulant expansion discussed in Chapter 3.

Bayes mixture model used in this chapter is remarkably *compact*; it requires a very small number of parameters relative to alternative forms of the distribution $p(\mathbf{x})$. In fact, for the binary alloy predictions presented here, a naive Bayes mixture model with $m = 9$ mixture components requires just 1.1% of the number of parameters required for the cumulant expansion [3] ! This dramatic reduction in the required number of parameters for roughly an equivalent predictive power indicates that latent class structure is an easy way to explain the appearance of crystal structure in binary alloys. In other words, the prototypes shared between two or more alloy systems define associations that can be utilized for prediction purposes [4].

[3] the pairwise cumulant expansion (Equation 3.8) requires one parameter for every pair probability

[4] in the same spirit as collaborative filtering systems [139]

## 4.3 Post-analysis of alloy classes

In Section 4.1 the use of a mixture model was motivated on the basis of reducing model complexity. Using the EM algorithm for parameter estimation and cross-validation to determine the optimal number of components, $m_{\text{opt}}$, we have now constructed a *compact* probabilistic model for binary alloy ground states with significant predictive power (as demonstrated in Figures 4-3 and 4-4). In Section 4.1 also discussed the underlying theme behind mixture models; *classification.* The idea is that classes or groups of alloys sharing similar prototypes drive the predictive power of mixture models. In a way, predictions in a mixture model are performed in a similar manner to the recommendations made by collaborative filtering [139] systems. For example, suppose the alloy A-B contains the structure prototypes $\alpha$ and $\beta$, while alloy C-D contains the prototypes $\alpha$ and $\gamma$. When presented with a new alloy, also containing the structure prototype $\alpha$, the information contained in alloys A-B and C-D can be leveraged to recommend $\beta$ and $\gamma$ as likely prototypes for the new alloy based on the fact that $\alpha$ is present. The mixture model presented in this chapter derives its predictive power through a related mechanism. The EM algorithm is used to group alloys together forming the class-conditional distributions $\{p(\mathbf{x}|j)\}$. These class-conditional distributions encode the information about *what* the groups of alloys are and the structure prototypes that are most likely to be shared within a group. In this section the mixture model fitted with the EM algorithm will be used to analyze *what* classes are present in the Pauling File binary alloy database and whether this latent class structure is physically meaningful. An alloy's class membership can be estimated by constructing a distribution over classes $j$ given the alloy's ground states, $\mathbf{x}$ – viz.

$$p(j|\mathbf{x}) = \frac{p(\mathbf{x}, j)}{p(\mathbf{x})} = \frac{p(j)\prod_i p(x_i|j)}{\sum_k p(k)\prod_i p(x_i|k)}$$

Note that the mapping from alloy ground states, $\mathbf{x}$, to class membership, $j$, is done in a *soft* way. In other words, $p(j|\mathbf{x})$ is the probability of alloy $\mathbf{x}$ belonging to class $j$, which can take on any value between 0 and 1 – hence the assignment of $\mathbf{x}$ to $j$ is soft. Because of these soft assignments, it is possible for an alloy to "belong"

to several different classes simultaneously. To back out the alloy classes identified by the EM fitting procedure we start by calculating the distribution $p(j|\mathbf{x})$ for each alloy in the database $\mathcal{D}$. The alloy is then placed into group $j$ if $p(j|\mathbf{x}) > (1-\epsilon)$ where $\epsilon \approx 10^{-3}$ – when this condition holds, the alloy is strongly predicted to belong to class $j$. After scanning over all alloys, the available data, $\mathcal{D}$, is now split into a set of alloy classes, $\mathcal{D}_1, \mathcal{D}_1, \ldots, \mathcal{D}_m$. Each alloy class, $\mathcal{D}_i$, should contain a collection of systems sharing similar structure prototypes – this is the mechanism by which alloys are classified. Because the alloys present in class $i$ (i.e., $\mathcal{D}_i$) will share common structure prototypes, each class should contain alloys that are chemically similar. To test the validity of this assertion we have plotted the binary alloy classes present in the Pauling File database in a Pettifor map [4], shown in Figure 4-5. In Figure 4-5 each $(x, y)$ coordinate corresponds to a binary alloy. Each element in the periodic table is given a unique number, called the Mendeleev number described in Ref. [4], and this number is used to define the coordinate system in Figure 4-5. Each alloy class, of which there are $m = 9$, is assigned a different symbol in Figure 4-5. Alloys in the database satisfying the condition $p(j|\mathbf{x}) > (1-\epsilon)$ with $\epsilon = 10^{-3}$ are plotted in Figure 4-5 according to their class membership. A key feature of Figure 4-5 is that alloy classes group together in the 2-d space defined by an element's Mendeleev number. Elements with similar Mendeleev numbers are often "similar" in a chemical sense (the Mendeleev scale tends to run up and down the columns of the periodic table). For example, the dark blue squares in Figure 4-5 correspond to alloy class $j = 2$. Each alloy in this class combines a rare-earth or group IIA or IIIA element in the periodic table *with* a late-transition metal series element such as Rh, Ir, Ni, or Pd. Consider also the alloy class given by $j = 4$ corresponding to the gray square symbols in Figure 4-5. This group consists of alloying Ge, Sn, Sb, Bi, or Po, all elements just to the *left* of the Zintl line in the periodic table, with just about anything else. The group $j = 0$ (red squares) corresponds to alloys containing one rare-earth element and one of Tl, In, Ga, or Al. Figure 4-5 confirms in a graphical way, the chemical intuition underlying the construction of the mixture model presented in this chapter. Note however, that no *assumptions* have been made to arrive at this grouping – these

are alloy classes determined on-the-fly based on the appearance of common structure prototypes. The mixture model derives its prediction power by grouping together similar alloys contained within the database $\mathcal{D}$. The groups of alloys determined by the EM algorithm are chemically similar, as evidenced by their spatial grouping when projected into the Pettifor map (i.e., neighboring elements on the Mendeleev scale are chemically similar, so clusters of points in the Pettifor map are indicative of chemical similarity). Similarity within each alloy group is determined by structure prototypes shared across alloys within the group. When presented with a new alloy for which only partial information is available, the mixture model predicts the most likely structures on the basis of how the partial information aligns with the presently known alloy groups.

## 4.4 Summary

In this chapter we have developed a probabilistic model for structure suggestion using a naive Bayes mixture model. Mixture models are used in a wide variety of machine learning problems from classifying hard drive failures [140] to understanding gene expression data [141]. Mixture models are useful for identifying latent class structure in data that can be used for prediction purposes. We have shown that a naive Bayes mixture model performs remarkably well in the task of predicting the ground states of binary alloys. It achieves a level of predictive power rivaling the cumulant expansion presented in Chapter 3, but does so with a dramatic reduction in the number of model parameters required. The latent class structure obtained for the Pauling File dataset indicates that a mixture model derives its predictive power by grouping together chemically similar alloys. In constrast to other techniques for alloy classification, such as the structure mapping technique presented in Section 1.2.2, the mixture model classification scheme utilizes the physics driving structure stability at many different compositions to defining class structure.

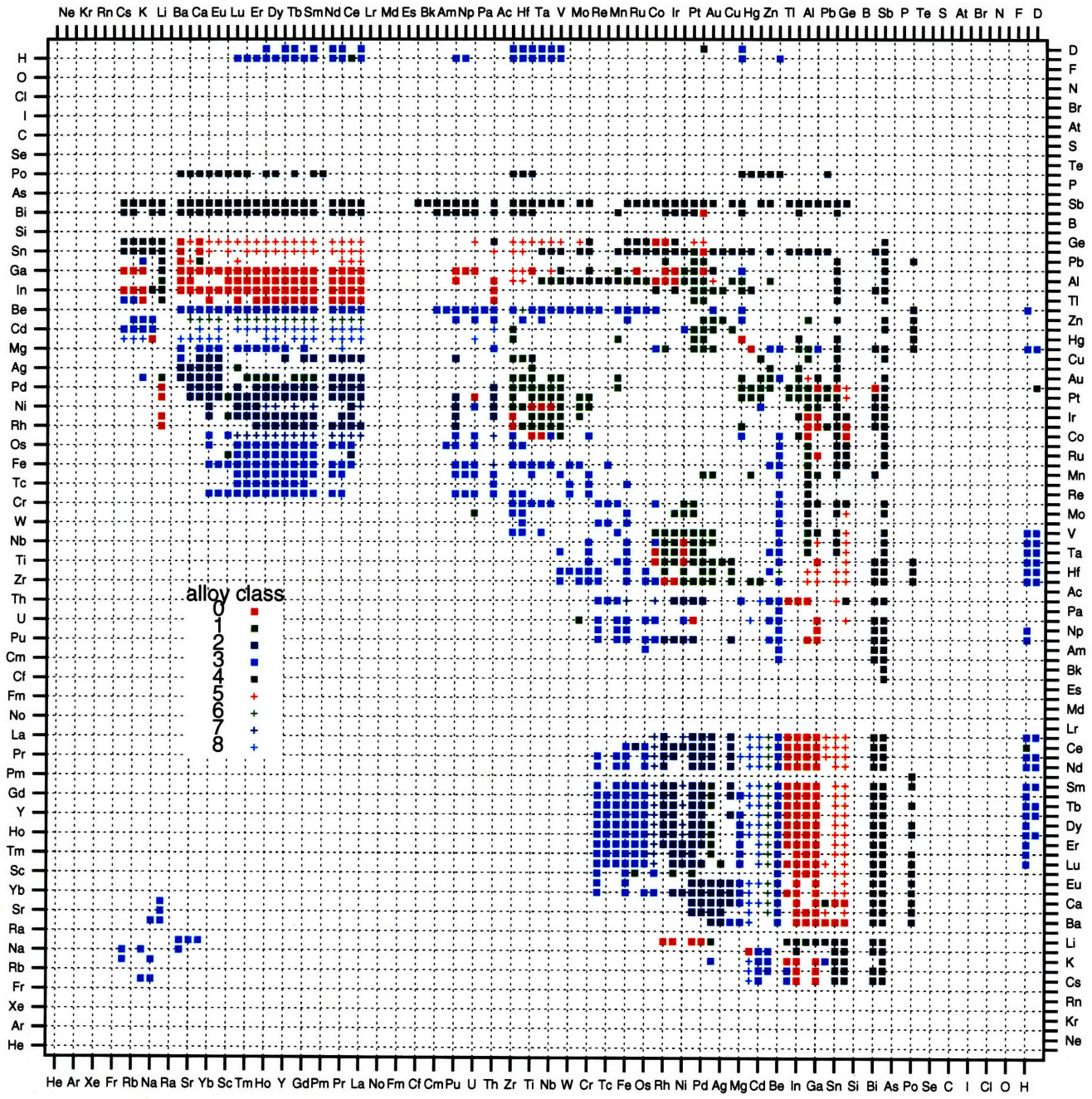


Figure 4-5: Pettifor map of alloy classes present in the Pauling File binaries edition database [2] for a mixture model with $m = 9$ components. Each symbol plotted represents an alloy for which $p(j|\mathbf{x}) > 0.999$. Elements are ordered on each axis according to their Mendeleev number described in Reference [4]. Alloys not containing "metallic" elements, as defined in Section 2.2, are deliberately ignored.

# Chapter 5

# Conclusions and future research

Computational materials science has made significant progress toward the goal of a virtual materials design laboratory. The exponential growth in FLOPS per dollar will continue to push outward the boundary of materials science problems that can be aided through calculation. As the field turns its eye toward problems requiring a search over chemistries and compositions, methods will be needed to suggest *what* to calculate where nothing is already known. In this thesis we have developed an abstract machine learning framework (Figure 1-2) which takes as input a collection of historical and computed data and as output furnishes the user with a list of highly informed suggestions for detailed investigation. Our method ultimately uses correlations present in a database of information, $\mathcal{D}$, along with available evidence, $\mathbf{e}$, to perform the prediction process. An investigation of specific structure correlations (Tables 2.1 and 2.2) provides evidence that our method is *consistent* with more traditional heuristic-based rationalizations of structure stability. To incorporate individual correlations into a coherent system for making predictions we have adopted the use of probabilistic graphical models. We have applied this technique to the study of binary metallic alloys, and have demonstrated through both proof-of-concept and statistical analysis, that it is highly *efficient* in making suggestions for further calculation.

Nevertheless, a number of open questions remain and suggestions for future research are given below.

## 5.1 Suggestions for future research

### 5.1.1 Applications

Perhaps the most fruitful direction for future research is to apply the methods developed in this thesis for predictions in both binary and multi-component systems. Although binary alloys have been studied quite extensively, our experience with the Ag-Mg, Au-Zr, and Li-Pt alloy systems suggests that a number of compounds have yet to be discovered. Thanks to the availability of computing power and robust ab initio electronic structure codes, it is likely that the structure of these as yet undiscovered compounds can be predicted in a reliable way faster than detailed experiments can be performed. The probabilistic models constructed in this thesis can be used to rapidly guide calculations towards the most likely set of candidate structures for unknown compounds. Therefore, one application of our framework would be to fill in these missing holes in binary alloy data through the use of quantum mechanical calculations. Quite simply, this would consist of generating predictions followed by calculations for all binary alloys and compositions where experimental information is lacking or inconsistent. In addition, there is little doubt that a number of structure mis-assignments have been made throughout the course of experimental history.

Multi-component systems comprise another application of the technique presented in this thesis. We believe the potential for impact is quite large in multi-component alloys as they have not been studied as extensively as binary systems. In principle, the only ingredient required for this task is a database of compounds identified by their structure prototype. A recently developed structure prototyping algorithm [142, 143] has made it possible to prototype an arbitrary database of crystal structure information and will certainly aid in performing this task. To perform predictions in multi-component systems, strategies are needed to managing the complexity involved as the number of chemical components is increased. The mixture model discussed in Chapter 4 was investigated with complexity reduction as a primary objective, so it may be particularly useful for making predictions in multi-component systems.

Although this thesis focused on predicting structure in binary metallic alloys, there are many other classes of technologically relevant chemistries. For example, multi-component oxides form a class of systems that are both physically interesting and technologically relevant. Predicting the structure and properties of $A_xB_yO_n$ oxides in which the A and B cations can take on multiple formal valence states is an outstanding problem for computational materials science. While this problem, at first glance, appears to be equivalent to predicting structure in a ternary system, it can be simplified by considering only pseudobinary mixtures of $A^{n+}$ and $B^{m+}$ cations where $n$ and $m$ are allowed valence states of the A and B cations.

### 5.1.2 Method development

#### discovering new prototypes

A number of questions remain which require method developments beyond those discussed in this thesis. A significant drawback of the machine learning framework presented in this thesis is an inability to suggest truly new prototype structures. The domain of our prediction ability is limited to all currently known structure prototypes and nothing further. While there is some evidence that the number of unknown structure prototypes is small (e.g., by analyzing the number of new prototypes discovered as a function of time), a true structure prediction scheme should have the ability to suggest entirely new structure prototypes. This drawback of our method is a result of the fact that we have explicitly avoided developing a microscopic description of a system's energetics with an *approximate* Hamiltonian. While approximate Hamiltonians are useful in many other areas of the materials design problem, we refrained from using them in an attempt to maximize the use of historical data. To discover truly new structure prototypes, a method is needed to intelligently explore the infinite space of unknown structures. The genetic algorithm (GA), a stochastic optimization method discussed in Section 1.2.1, appears to provide a intriguing combination of general applicability (i.e., it isn't restricted to any particular *class* of systems) and relative

efficiency (i.e., relative to the only other stochastic technique, simulated annealing). However, we believe current implementations of the GA are notably inefficient. The literature available on the subject indicates that a very large number of total energy evaluations and relaxation steps must be performed before convergence can be obtained (as discussed in Section 1.2.1). Current implementations of the GA are initialized with a population of structures with *random* coordinates, forcing the method to spend a considerable amount of effort optimizing away from energetically unfavorable configurations. We believe a particularly simple way to improve the efficiency of the GA would be to *seed* the initial population with a set of highly likely structures suggested with a machine learning model. Doing so will initialize the population of structures in energetically *relevant* portions of phase space. The algorithm then proceeds forward with mating and mutation steps providing the stochastic component required to discover new prototypes.

**pruning the cumulant expansion**

The cumulant expansion discussed in Chapter 3 performed remarkably well in predicting crystal structure. In Chapter 3 correlation terms, $g_{ij}(x_i, x_j)$, were included for *all* distinct pairs of variables. While the correlation terms gave rise to significant prediction improvements, it is possible that some terms simply lead to over-fitting (reducing the prediction ability). No systematic procedure was formulated for deciding which correlation terms to include. Therefore, one possibility for increasing the prediction capability of Equation 3.8 is to remove correlation terms which decrease the prediction performance. For example, one could start with the independent variable approximation (Equation 2.4) and include correlation terms through a greedy search (i.e., pick the correlation term which improves the prediction performance the most, then the next, and so on). Another option would be to start with full pairwise expansion (Equation 3.8) and remove the correlation term which decreases the prediction performance the most (if present at all).

### fitting maximum entropy models

The maximum entropy model presented in Section 3.3.2 resolved the formal problems associated with a cumulant expansion for $p(\mathbf{x}|\mathcal{D})$. The maximum entropy solution was not pursued further due to the difficulty in solving for the Lagrangian multipliers, $\boldsymbol{\lambda}$. Although not discussed in this thesis, a preliminary investigation of solving for the Lagrangian multipliers has been performed on several models containing three and four variables (so that the expectation step could be performed analytically). To perform the expectation step on the full model approximate methods must be pursued and Monte Carlo sampling has been investigated. However, the very large number of equilibration and sampling steps were required to obtain reasonable estimates of the marginal probabilities.

Several alternative techniques for performing approximate expectation have been recently developed in the machine learning community [100, 127, 128]. Oddly enough, these approximate techniques are equivalent [127] to Kikuchi's CVM formalism [102, 101] for constructing approximate free energy functionals. There is some evidence that these techniques approximate the expectation step appearing in Algorithm 2 more accurately than Monte Carlo sampling [129]. Therefore, if one were to develop the maximum entropy approach further, a fruitful direction would be to utilize the CVM for performing the expectation step in Algorithm 2. Moving beyond the traditional CVM, a technique in active development by Wainwright, Jaakkola, and Willsky [128] expresses a CVM-like entropy as a convex combination of entropies that can be exactly computed. Doing so results in an algorithm with better stability properties than the original CVM formalism.

# Appendix A

# Notation, Probability, and related functions

Although it may seem a little silly at first glance, this appendix is devoted to an overview of the notation used in this thesis when talking about probabilities. The decision to do so is based on the observation that there is simply no consistent notation in the theory of probability. Every author seemingly uses their own notation, leading to a confusing state of affairs for the reader. Thus for the sake of readability, a notational overview and review of the basic properties of probability functions is given here. This appendix borrows heavily from Chapter 2 of MacKay's book *Information Theory, Inference, and Learning Algorithms* [96].

## A.1 random variables and their probabilities

> Just as the White Knight distinguished between the song, the name of the song, and what the name of the song was called, we will sometimes need to be careful to distinguish between a random variable, the value of the random variable, and the proposition that asserts the random variable has a particular value. –David MacKay [96]

For our purposes, a capitalized symbol will denote a random variable. A lower-case symbol will denote the outcome of a random variable which is one element in a set

of possible values. For example, $X$ will denote a random variable, $x$ the outcome of a random variable, and $\Omega_X \equiv \{v_1, v_2, \ldots, v_{q_X}\}$ a set of $q_X$ possible outcomes. For every possible outcome of $X$ we will associate a probability, $p_i$ in the set $\mathcal{P}_X = \{p_1, p_2, \ldots, p_{q_X}\}$ with $p(x = v_i) = p_i$ and $\sum_{v_i \in \Omega_X} p(x = v_i) = 1$. When the context is clear, a more compact notation will often be used such that $p(x = v_i)$ might be written as $p(v_i)$ or just $p(x)$. So the above normalization condition may be written a little more succinctly as $\sum_x p(x) = 1$.

## A.2 basic properties

### A.2.1 Joint probabilities

When two or more random variables are involved, say $X$ and $Y$, the possible outcomes of both are identified with an ordered pair $x, y$ where $x \in \Omega_X = \{v_1, \ldots, v_{q_x}\}$ and $y \in \Omega_Y = \{w_1, \ldots, w_{q_Y}\}$. The set of all possible outcomes of $X$ and $Y$ is just the Cartesian product over the domains of each variable or $\Omega_{X,Y} = \Omega_X \times \Omega_Y$. The number of elements in the set of possible outcomes is $q_X * q_Y = |\Omega_X||\Omega_Y|$. We will let $p(x, y)$ denote a joint probability, i.e., the probability of an event where $X$ and $Y$ take on the values $x$ and $y$ respectively. Continuing in this fashion, if one has the set of variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ the probability of event $\mathbf{x} \in \Omega_{X_1} \times \Omega_{X_2} \times \cdots \times \Omega_{X_n} \equiv \Omega_{\mathbf{X}}$ is denoted $p(\mathbf{x})$ where $|\Omega_{\mathbf{X}}| = \prod_i q_i$. Subject only to the normalization condition $\sum_{\mathbf{x}} p(\mathbf{x}) = 1$ there are $|\Omega_{\mathbf{X}}| - 1$ independent values of $p(\mathbf{x})$.

### A.2.2 Marginalization

From $p(x, y)$ we can obtain the marginal probability, $p(x)$, through summation

$$p(x) = \sum_y p(x, y) \tag{A.1}$$

### A.2.3 Conditional probability

The probability of $X$ having outcome $x$ given that $Y$ has the outcome $y$ is given by

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(x,y)}{\sum_{x'} p(x',y)} \tag{A.2}$$

For the above equation to have meaning we must obviously have $p(y) > 0$. Conditional probabilities provide a mechanism through which the probabilities over $x \in \Omega_X$ can be updated provided knowledge of the outcome of $Y$.

### A.2.4 Product rule

The product rule is obtained from the definition of a conditional probability

$$p(x,y) = p(x|y)p(y) = p(y|x)p(x) \tag{A.3}$$

The last equality is also known as Bayes' theorem. Another equivalent form of the product rule is

$$p(x,y|\mathcal{I}) = p(x|y,\mathcal{I})p(y|\mathcal{I}) \tag{A.4}$$

where $\mathcal{I}$ is any other information.

### A.2.5 Independence

Two variables $X$ and $Y$ are independent iff

$$p(x,y) = p(x)p(y) \tag{A.5}$$

so that

$$p(x|y) = \frac{p(x,y)}{p(y)} = p(x) \tag{A.6}$$

i.e., the fact that the outcome of $Y$ is $y$ has no effect on the probability assigned to the outcome $x$ of $X$. A mathematical statement of this independence is denoted $(X \perp Y)$. When two variables $X$ and $Y$ are independent given a third variable, say

$Z$, we denote the independence statement $(X \perp Y|Z)$ and we must have

$$p(x, y|z) = p(x|z)p(y|z) \tag{A.7}$$

## A.3 Information entropy and related functions

Further details regarding Shannon's information entropy [120] and its properties are summarized in the paper [121] and book by Jaynes [122]. MacKay [96] and the book by Cover and Thomas [91] also give a thorough review. Some basic properties are given below.

The information entropy associated with the random variable $X$, under the distribution $p(x)$, denoted $H(X)$, $H_X$, or $H[p(x)]$ is given by

$$H(X) \equiv -\sum_x p(x) \log(p(x)) \tag{A.8}$$

Because $0 \leq p(x) \leq 1$, the information entropy $H(X) \geq 0$ with equality iff $p(x) = 1$ for some $x \in \Omega_X$. Entropy is a maximum if $p(x)$ is a uniform distribution or $p(x) = |\Omega_X|^{-1}$ in which case $H(X) = \log(|\Omega_X|)$. For distributions defined over more than one variable, say $X$ and $Y$, the information entropy is given by

$$H(X, Y) = H_{X,Y} = -\sum_{x,y} p(x, y) \ln(p(x, y)) \tag{A.9}$$

When the variables are independent, the entropy is additive

$$\begin{aligned} H(X, Y) &= -\sum_{x,y} p(x)p(y) \left(\ln(p(x)) + \ln(p(y))\right) \\ &= -\sum_x p(x) \ln(p(x)) - \sum_y p(y) \ln(p(y)) \\ &= H(X) + H(Y) \end{aligned} \tag{A.10}$$

Note the connection to thermodynamics where entropy (that is thermodynamic entropy) is defined as an extensive or additive quantity. When extensivity holds, a system consisting of two identical subsystems, each with $N$ particles, will have a probability distribution that decomposes as independent distributions for each of the $N$-particle systems.

$$p(x_1, x_2, \ldots, x_{2N}) = p(x_1, \ldots, x_N)p(x_{N+1}, \ldots, x_{2N})$$

so

$$H(X_1, \ldots, X_{2N}) = 2H(X_1, \ldots, X_N)$$

Using the product rule (Equation A.3) we can express the entropy as

$$\begin{aligned} H_{X,Y} &= -\sum_{x,y} p(x,y) \ln (p(x,y)) \\ &= -\sum_{x,y} p(x,y) \left[\ln (p(x)) + \ln (p(y|x))\right] \\ &= H_X - \sum_{x,y} p(x,y) \ln (p(y|x)) \\ &= H_X + H_{Y|X} \end{aligned} \tag{A.11}$$

Equation A.11 defines the *conditional entropy*, $H_{Y|X}$. Using successive applications of the product rule one can show that for the variables $X_1, \ldots, X_n$

$$H_{\mathbf{X}} = \sum_{i=1}^{n} H_{X_i|X_1,\ldots,X_{i-1}}$$

### A.3.1 Kullback-Leiber divergence

The relative entropy or Kullback-Leiber divergence between two probability distributions $p(x)$ and $q(x)$ is given by

$$D_{KL}(p||q) = \sum_x p(x) \log \left(\frac{p(x)}{q(x)}\right) \tag{A.12}$$

Although not obvious, $D_{KL}(p||q) \geq 0$ with equality iff $p(x) = q(x)\ \forall\ x \in \Omega_X$. The statement $D_{KL}(p||q) \geq 0$ can be proved using Jensen's inequality [91]. The Kullback-Leiber divergence measures the "distance" between the distributions $p(x)$ and $q(x)$, although it cannot be interpreted as a distance metric because $D_{KL}(p||q) \neq D_{KL}(q||p)$. Nevertheless $D_{KL}(p||q)$ is useful for say measuring overall difference between the distribution $p(x)$ and a target distribution $q(x)$. For example, $D_{KL}(p||q)$ can be used to derive the variational principle of statistical mechanics [32].

### A.3.2 Mutual information

The mutual information between variables $X$ and $Y$ is given by the Kullback-Leiber divergence between $p(x, y)$ and the distribution $q(x, y) = p(x)p(y)$

$$I(X;Y) = I_{X,Y} = \sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

The mutual information measures the degree to which two variables are correlated, or equivalently, how much the outcome of one variable constrains the possible outcomes of the other. The mutual information $I_{X,Y}$ is symmetric $I_{X,Y} = I_{Y,X}$ and bounded by $0 \leq I_{X,Y} \leq \min(H_X, H_Y)$. The equality $I_{X,Y} = 0$ is obtained iff the two variables are independent or uncorrelated, while $I_{X,Y} = \min(H_X, H_Y)$ is obtained only if one of the variables is a deterministic function of the other.

**proof of $0 \leq I_{X,Y} \leq \min(H_X, H_Y)$**

Because $D_{KL}(q||p) \geq 0$, we know that $I_{X,Y} = D_{KL}(p(x,y)||p(x)p(y)) \geq 0$. Using definition of conditional entropy

$$\begin{aligned} I_{X,Y} &= H_X + H_Y - H_{X,Y} \\ &= H_Y - H_{Y|X} \\ &= H_X - H_{X|Y} \end{aligned}$$

Using the inequalities $0 \leq H_{Y|X} \leq H_Y$ we can re-arrange terms to arrive at

$$0 \leq I_{X,Y} \leq H_Y$$

Using the inequalities $0 \leq H_{X|Y} \leq H_Y$ gives

$$0 \leq I_{X,Y} \leq H_X$$

Hence $0 \leq I_{X,Y} \leq \min(H_X, H_Y)$.

# Appendix B

# Parameter Estimation

To use a probabilistic model of structural stability we have to at some point connect a database of information, $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, into the numerical values assigned to the function $p(\mathbf{x})$. The values of $p(\mathbf{x})$ are determined through a collection of parameters, here denoted $\boldsymbol{\theta}$. Many values of $\boldsymbol{\theta}$ are permissible, but at the end of the day we must make a choice, and parameter estimation is the process through which the choice is made. For example, suppose you've been tossing a biased coin and you're given the question, "what is the probability that the next toss is heads ?" Any answer between 0 and 1 is permissible, and this section outlines two different strategies used to obtain an answer. Before detailing parameter estimation, it is important to note the following. One must first pick the form for a probability distribution and this will determine both the number and semantics of the parameters used. For example, to analyze some univariate ordinal data one could use either a Poisson distribution, with just one parameter, or a Normal distribution, with two. Choosing the form of a distribution, called *model selection*, is not described here. Rather, parameter estimation is the process of choosing a particular $\boldsymbol{\theta}$ from the infinite set of possibilities *for a fixed model.* We will also restrict the scope of this Appendix to closed-form parameter estimation. There are models for which it is not possible to write down a closed form solution for their parameters given some data and these will not be discussed here (e.g., undirected graphical models described in Chapter 2). None of this material is "new", but is provided as support for Chapters 2 and 3.

## B.1 multinomial

The problems discussed in this thesis deal with discrete-valued variables containing a finite number of possible outcomes. The multinomial distribution is appropriate for such cases, so multinomial parameter estimation is described here. Consider a variable $X$ that can take on $q$ possible values, i.e., $x \in \{v_1, v_2, \ldots, v_q\}$. Given a set of N observations for this variable, denoted $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ we are interested in determining the probability function $p(x|\mathcal{D})$. For each of the $q$ possible values of $X$ we will associate a parameter with the *value* of the probability function. In other words, let $\theta_{v_i}$ denote the parameter associated with the $i^{th}$ value of $X$ or $p(x = v_i) = \theta_{v_i}$. Because we are dealing with a multinomial distribution the parameters **are** the values assigned to the function $p(x)$ – in contrast to say a Normal distribution where you would estimate the mean ($\mu$) and variance ($\sigma^2$) from available data and use the formula $p(x) \propto \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$ to determine probability *values*. The collection of $q$ parameters will be referred to as $\boldsymbol{\theta} = \{\theta_{v_i}\}_{i=1,\ldots,q}$. Given $\mathcal{D}$ there are several ways of choosing a particular $\boldsymbol{\theta}$ from the set of permissible values $\Theta \equiv \{\boldsymbol{\theta} : \sum_{i=1}^{q} \theta_{v_i} = 1\}$ discussed next.

### B.1.1 Maximum Likelihood

An approach often used is to choose $\boldsymbol{\theta}$ such that it maximizes the likelihood of observing the data given. For a particular value of $\boldsymbol{\theta}$, the log-likelihood of the data is given by

$$\begin{aligned} l(\mathcal{D};\boldsymbol{\theta}) &= \log p(\mathcal{D}|\boldsymbol{\theta}) \\ &= \log p(x_1, x_2, \ldots, x_N|\boldsymbol{\theta}) \\ &= \sum_{t=1}^{N} \log p(x_t|\boldsymbol{\theta}) \\ &= \sum_{t=1}^{N}\sum_{i=1}^{q} \delta(x_t, v_i) \log \theta_{v_i} \\ &= \sum_{x} n(x) \log \theta_x \end{aligned} \tag{B.1}$$

Where $\delta(x_t, v_i) = 1$ if $x_t = v_i$ and is zero otherwise, and $n(x)$ is the number of times that the variable $X$ had the outcome $x$ in the dataset. Maximizing the log-likelihood function under the constraint $\sum_x \theta_x = 1$ leads to the well-known maximum likelihood estimate for $\boldsymbol{\theta}$

$$\hat{\theta}_x^{ML} = \frac{n(x)}{\sum_{x'} n(x')} \text{ for } x \in \{v_1, \ldots, v_q\} \tag{B.2}$$

Here the optimal setting of $\boldsymbol{\theta}$, denoted by the hat symbol $\hat{\boldsymbol{\theta}}$, is given by emprical frequencies of observation. The counts $\{n(x)\}$ are known as *sufficient statistics* since they are sufficient to parameterize the distribution $p(x)$ from available data.

### B.1.2 Bayesian estimate

The central idea of a Bayesian estimate for $\boldsymbol{\theta}$ is to consider all permissible values of $\boldsymbol{\theta}$ as valid, but weigh them according to what the data would indicate. For example, suppose you are tossing a bent coin (or tack for that matter) and would like to measure the geometry of the bend (or length to width ratio of the tack) without the use of a micrometer (i.e., using only the laws of Newtonian mechanics and the outcomes of the tosses). After tossing the coin $N = 2000$ times, your estimate of the coin's geometry is much more strongly peaked around a "best guess" than after just $N = 2$ tosses. Bayesian parameter estimates are an attempt to capture this effect of various *degrees of belief* [122, 144, 96] when picking a particular value of $\boldsymbol{\theta}$. A Bayesian estimate will capture the uncertainty associated with $\boldsymbol{\theta}$ by constructing a distribution over the permissible values of $\boldsymbol{\theta}$ conditioned on the available data. The key point is that a Bayesian approach will incorporate both a "best guess" and our uncertainty in the "best guess" on the same footing. Additionally, a Bayesian framework allows statements to be made about what values of $\boldsymbol{\theta}$ are "sensible" before observing a single shred of data – i.e., using only information about the problem at hand. As a consequence of this latter property, a Bayesian approach allows for the possibility of assigning a probability values in data-limited settings (i.e. where the maximum likelihood estimate would assign a probability of zero to an event, although paradoxically such an event is permissible). This property of Bayesian estimates is

crucial for complex graphical models where the number of parameters in the model can approach the size of the available dataset. We start by expressing $p(x|\mathcal{D})$ as an integral over all possible values of $\boldsymbol{\theta}$ using the marginalization property of probabilities (Equation A.1)

$$p(x|\mathcal{D}) = \int p(x, \boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{B.3}$$

Using the product rule (Equation A.3) we can decompose this further

$$p(x|\mathcal{D}) = \int p(x|\boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{B.4}$$

Using the definition of a parameter ($p(x|\boldsymbol{\theta}, \mathcal{D}) = \theta_x$), we are left with the central object of interest: $p(\boldsymbol{\theta}|\mathcal{D})$. Applying Bayes' Rule we have

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &= p(\mathcal{D}|\boldsymbol{\theta}) \frac{p(\boldsymbol{\theta})}{p(\mathcal{D})} \\ &= p(x_1, x_2, \ldots, x_N|\boldsymbol{\theta}) \frac{p(\boldsymbol{\theta})}{p(x_1, \ldots, x_N)} \\ &= \lambda \left[ \prod_x \theta_x^{n(x)} \right] p(\boldsymbol{\theta}) \end{aligned} \tag{B.5}$$

Here the quantity $p(x_1, \ldots, x_N)$ has been absorbed into an overall normalization constant $\lambda^{-1} = p(x_1, \ldots, x_N)$ which will be determined using the constraint

$$1 = \int p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{B.6}$$

The quantity $p(\boldsymbol{\theta})$ represents a prior over our distribution's parameters. It represents our belief in how probabilities should be assigned based only on what is known about the problem at hand *before* receiving any data [122]. A convenient, consistent, and arguably the only [145] distribution for $p(\boldsymbol{\theta})$ in the multinomial case is the so-called Dirichlet distribution

$$p(\boldsymbol{\theta}) = \beta(\boldsymbol{\alpha}) \prod_x {\theta_x}^{\alpha_x - 1}$$

where $\beta(\boldsymbol{\alpha}) = \frac{\Gamma(\sum_x \alpha_x)}{\prod_x \Gamma(\alpha_x)}$. A Dirichlet distribution is *conjugate* to a multinomial distribution in the sense that when you use a Dirichlet prior with your data, you obtain a multinomial *posterior* distribution over the permissible parameter values. So using a Dirichlet we are left with

$$p(\boldsymbol{\theta}|\mathcal{D}) = \lambda\beta(\alpha)\prod_x {\theta_x}^{n(x)+\alpha_x-1} \tag{B.7}$$

To determine $\lambda$ we will need to perform the following integral under the constraint $\sum_x \theta_x = 1$

$$I(\mathbf{n}, \boldsymbol{\alpha}) = \int \prod_x \theta_x^{n(x)+\alpha_x-1} d\theta_{v_1} d\theta_{v_2} \cdots d\theta_{v_q} \tag{B.8}$$

where $\mathbf{n} = \{n(v_1), \ldots, n(v_q)\}$. To enforce the constraint $(\sum_x \theta_x = 1)$ we extend the range of integration for each component to $\theta_x \in [0, \infty)$ and make use of a delta function

$$I(\mathbf{n}, \boldsymbol{\alpha}) = \int_0^\infty \prod_x \theta_x^{n(x)+\alpha_x-1} \delta\left(\sum_{x'} \theta_{x'} - r\right) d\boldsymbol{\theta} \tag{B.9}$$

Laplace transforming both sides we have

$$\begin{aligned}
\frac{I(\mathbf{n}, \boldsymbol{\alpha})}{s} &= \int_0^\infty \int_0^\infty \prod_x \theta_x^{n(x)+\alpha_x-1} \delta\left(\sum_{x'} \theta_{x'} - r\right) \exp(-rs) dr d\boldsymbol{\theta} \\
&= \left(\int_0^\infty d\theta_{v_1} \theta_{v_1}^{n(v_1)+\alpha_{v_1}-1} \exp(-s\theta_{v_1})\right) \cdots \left(\int_0^\infty d\theta_{v_q} \theta_{v_q}^{n(v_q)+\alpha_{v_q}-1} \exp(-s\theta_{v_q})\right)
\end{aligned}$$

Using the definition of the gamma function $\left(\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt\right)$ and a change of variables we are left with

$$\frac{I(\mathbf{n}, \boldsymbol{\alpha})}{s} = \prod_x \frac{\Gamma\left(n(x) + \alpha_x\right)}{s^{n(x)+\alpha_x}}$$

Inverse transforming the above and evaluating at $r = 1$ yields

$$I(\mathbf{n}, \boldsymbol{\alpha}) = \frac{\prod_x \Gamma\left(n(x) + \alpha_x\right)}{\Gamma\left(\sum_{x'} n(x') + \alpha_{x'}\right)}$$

Using the normalization condition to determine $\lambda$ we are left with

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &= \Gamma\left(\sum_{x'} n(x') + \alpha_{x'}\right) \prod_{x} \frac{{\theta_x}^{n(x)+\alpha_x-1}}{\Gamma(n(x)+\alpha_x)} \\ &= C(\mathbf{n}, \boldsymbol{\alpha}) \prod_{x'} {\theta_{x'}}^{n(x')+\alpha_{x'}} \end{aligned} \tag{B.10}$$

Before moving on to finish the calculation of $p(x|\mathcal{D})$ we should perhaps say a few words about what $p(\boldsymbol{\theta}|\mathcal{D})$ gives us. First, we now have an object that encodes our various degrees of belief in the possible values of $\boldsymbol{\theta}$ both before and after a set of data, $\mathcal{D}$, is observed. The values of $\boldsymbol{\alpha} \equiv \{\alpha_{v_1}, \ldots, \alpha_{v_q}\}$ communicate our degrees of belief before any data is observed and these are updated according to what the observations indicate through the empirical counts $\mathbf{n}$. Taking this back to the bent coin example, if one knows the coin contains no bend, then it is sensible to choose very large and equal numbers for the two prior parameters $\alpha_{\text{heads}}$ and $\alpha_{\text{tails}}$. Picking such values for $\alpha_{\text{heads}}$ and $\alpha_{\text{tails}}$ leads to a prior distribution, $p(\theta_{\text{heads}}, \theta_{\text{tails}})$ that is very strongly peaked around $\theta_{\text{heads}} = \theta_{\text{tails}} = 1/2$. However, knowing the coin is bent removes a lifetime of prior knowledge and a more sensible prior would correspond to an $\alpha_{\text{heads}}$ and $\alpha_{\text{tails}}$ that are much smaller.

Getting back to the program, we are interested in $p(x|\mathcal{D})$ when $x$ is equal to some value $x \in \{v_1, \ldots, v_q\}$.

$$\begin{aligned} p(x = v_i|\mathcal{D}) &= \int \theta_{v_i} p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \\ &= C(\mathbf{n}, \boldsymbol{\alpha}) \int \theta_{v_i} \prod_{x'} {\theta_{x'}}^{n(x')+\alpha_{x'}} d\boldsymbol{\theta} \\ &= C(\mathbf{n}, \boldsymbol{\alpha}) I(\{n(v_1), \ldots, n(v_i)+1, \ldots\}, \boldsymbol{\alpha}) \\ &= \frac{n(v_i) + \alpha_{v_i}}{\sum_{x'} n(x') + \alpha_{x'}} \end{aligned} \tag{B.11}$$

This result is intuitively consistent with the view that $\boldsymbol{\alpha}$ communicates a prior belief and the values of $\alpha_x$ appear simply as extra "counts" in the problem. One can also view the parameters $\boldsymbol{\alpha}$ as "smoothing" parameters; they lower the variance of

parameter estimates over different datasets. Due to their formal properties and well-known advantages [95], Bayesian parameter estimates are used throughout this thesis. Empirical tests comparing Bayesian parameter estimates to Maximum Likelihood estimages often show that Bayesian estimates more closely approximate the true distribution for a finite amount of data, $\mathcal{D} = \{x_1, \ldots, x_N\}$ [95]. Note that $p(x|\mathcal{D})$ is given by the expection value of $\theta_x$ over the distribution $p(\boldsymbol{\theta}|\mathcal{D})$ denoted $\langle\theta_x\rangle$. With this in mind, the uncertainty in $p(x|\mathcal{D})$ can be related to

$$\sigma_x{}^2 = \langle\theta_x{}^2\rangle - \langle\theta_x\rangle^2$$

For reference, the $m^{th}$ moment of $\theta_x$ under the distribution $p(\boldsymbol{\theta}|\mathcal{D})$ is

$$\langle\theta_{v_i}{}^m\rangle = \frac{\Gamma\left(\sum_x n(x) + \alpha_x\right)}{\Gamma\left(m + \sum_x n(x) + \alpha_x\right)} \frac{\Gamma(n(v_i) + \alpha_{v_i} + m)}{\Gamma(n(v_i) + \alpha_{v_i})}$$

### B.1.3 multiple variables

When several variables are involved, say $X$ and $Y$, estimates are needed for the quantity $p(x, y|\mathcal{D})$ where $\mathcal{D}$ represents a set of $N$ observations of the ordered pairs $x, y$ or $\mathcal{D} \equiv \{(x, y)_1, (x, y)_2, \ldots, (x, y)_N\}$. If there are $q$ possible values of the variable $X$ and $r$ possible values of the variable $Y$, then there are $qr$ possible combinations of the pair. This situation is equivalent to the single variable case, except estimates are being performed for a multinomial with $qr$ possible values rather than just $q$ or $r$. In this case the Maximum Likelihood estimate will be given by

$$\hat{\theta}^{ML}_{x,y} = \frac{n(x, y)}{N}$$

and the Bayesian estimate

$$p(x = v_i, y = w_j|\mathcal{D}) = \frac{n(v_i, w_j) + \alpha_{v_i,w_j}}{N + \alpha_T}$$

Where $\alpha_T = \sum_{x,y} \alpha_{x,y}$.

### B.1.4 choosing an appropriate $\alpha$

In choosing a particular $\boldsymbol{\alpha}$ one is communicating a prior belief about the likelihood of possible outcomes before any data is observed. Thus no "standardized" mechanism exists for choosing $\boldsymbol{\alpha}$, rather the choice made will be somewhat problem dependent. Nevertheless, several common choices of $\boldsymbol{\alpha}$ exist in the literature [122, 136, 145] and it is useful to elaborate on the particular statement that each makes. One commonly used prior is specified by $\boldsymbol{\alpha} = (1, 1, 1, \ldots)$ which leads to a uniform distribution

$$p(\boldsymbol{\theta}) = \Gamma(q)$$

Such a prior distribution assigns a uniform probability to all possible parameter values and leads to parameter estimates that correspond to "Laplace's rule of succession" [122] or

$$p(x|\mathcal{D}) = \frac{n(x)+1}{N+q}$$

It can be seen that the above setting of $\boldsymbol{\alpha}$ has the effect of adding $q$ counts to the data, one observation for each possible value of the random variable $X$. In many settings one is interested in comparing the performance of different models. For example, one could choose to express $p(x, y)$ as a fully parameterized joint distribution, or use the approximation $p(x, y) = p(x)p(y)$. Using a uniform prior in each case will lead to $qr$ additional counts in the fully paramerized case and $q + r$ in the independent variable approximation. Having too many "additional" counts can lead to situations where the prior will overtake the effect of the data making it impossible to select one model over another. Thus, a more common choice for $\boldsymbol{\alpha}$ is

$$\alpha_x = \frac{n'}{q}$$

for the point estimate and

$$\alpha_{x,y} = \frac{n'}{qr}$$

for the joint probability estimate. This will have the effect of adding $n'$ counts to each estimated function. When $n' = 1$, the resulting prior corresponds to the so-called Minimum Information Dirichlet prior [136] which is used throughout this thesis.

# Appendix C

# DFT Calculations

All of the prediction results presented in this thesis would not be possible without an *accurate* method for evaluating the total energies of phases competing for stability. Our approach is to feed the predictions of a machine learning method into an accurate Hamiltonian. Doing so leverages the suggestive character of a machine learning technique with the accuracy of detailed quantum mechanical calculations (which are, at present, computationally expensive). At the time of writing this thesis, the most accurate and practical method for evaluating the total energy of a system is through the use of Density Functional Theory (DFT) [26]. Density functional theory has proven highly accurate in reproducing a wide range of materials properties [5]. In particular, significant agreement between DFT and experiment with regard to the stability of competing crystal structures has been shown by Curtarolo, Morgan, and Ceder [3]. Other total energy methods, such as Quantum Monte Carlo [146], the GW approximation [147, 30], and Dynamical Mean Field Theory [31] are available but presently only at a significant computational cost. Most importantly, a modern, robust implementation of the algorithms needed to perform DFT calculations is available in the Vienna Ab-Initio Simulation Package (VASP) [148]. This robust, thorougly tested implementation of DFT has enabled a large number of calculations to be performed with ease. This section gives a description of the most important parameters controlling the convergence of DFT calculations performed with VASP – it is included for posterity.

**$E_{xc}$ functional** Calculations performed in the Ag-Mg and Au-Zr alloy systems used the Perdew-Wang GGA exchange correlation functional (PW91) [149] while those performed in the Li-Pt system used the Perdew-Burke-Ernzherof GGA functional [150]. All calculations were performed spin-polarized.

**pseudopotentials** Projector Augmented Wave (PAW) [151] potentials were used throughout this thesis. The potentials were obtained from an all-electron calculation of a neutral isolated atom using a method described in by Kresse and Joubert[152]. Electrons treated as valence states (out of the core) for each element were as follows: Ag($4d^{10}$ $5s^1$), Mg($2p^6$ $3s^2$), Au($5d^{10}$ $6s^2$), Zr($4s^2$ $4p^6$ $5s^1$ $4d^3$), Li($2s^1$), Pt($5d^9$ $6s^1$)

**cutoff energy** Each pseudopotential requires a characteristic number of Fourier components required to describe the variations of its valence wavefunctions and pseudo core electron potentials accurately. This concept has been used to derive a suggested energy cutoff [153] for a plane wave basis set. For all calculations presented in this thesis energy cutoffs were set to 1.5 times the largest suggested cutoff for any species present in an alloy.

**k-point grid** Brillouin-zone integrations were performed using a Monkhorst-Pack mesh [154] containing at least 2500/(number of atoms in unit cell) **k**-points distributed as uniformly as possible over the reciprocal cell.

**coordinate optimization** For each predicted structure, the coordinates of the system are initialized to those of the structure prototype and a conjugate-gradient based optimization is performed over cell lattice vectors and positions of all atoms in the system.

**extra notes** all calculations are performed at zero temperature and pressure and the zero-point motion of the nuclei is neglected. Total energies are expected to be converged to $\approx 10\frac{\text{meV}}{\text{atom}}$ [3] while formation enthalpies, $\Delta H_{\text{form}}(A_{1-x}B_x) = H(A_{1-x}B_x) - (1-x)H(A) - xH(B)$, are expected to be converged to a much smaller tolerance due to cancellation of errors. These convergence values are

only up to the intrinsic approximations made in DFT, the particular GGA functional used, and the use of frozen-core PAW potentials.

# Bibliography


[1] T.B. Massalski, editor. *Binary Alloy Phase Diagrams.* American Society for Metals, 1987.

[2] P. Villars, M. Berndt, K. Brandenburg, K. Cenzual, J. Daams, F. Hulliger, T. Massalski, H. Okamoto, K. Osaki, A. Prince, H. Putz, and S. Iwata. The pauling file, binaries edition. *J. of Alloys and Compounds*, 367:293–297, 2004.

[3] S. Curtarolo, D. Morgan, and G. Ceder. Accuracy of *ab initio* methods in predicting the crystal structures of metals: A review of 80 binary alloys. *Computer Coupling of Phase Diagrams and Thermochemistry*, 29:163–211, 2005.

[4] D. G. Pettifor. The structures of binary compounds: I. phenomenological structure maps. *Journal of Physics C: Solid State Physics*, 19:285–313, 1986.

[5] J. Hafner, C. Wolverton, and G. Ceder. Toward computational materials design: the impact of density functional theory on materials research. *MRS Bulletin*, 31:659–668, 2006.

[6] Y. Wang, S. Curtarolo, C. Jiang, R. Arroyave, T. Wang, G. Ceder, L. Q. Chen, and Z. K. Liu. Ab initio lattice stability in comparison with calphad lattice stability. *Computer Coupling of Phase Diagrams and Thermochemistry*, 28:79–90, 2004.

[7] N. Marzari. Realistic modeling of nanostructures using density functional theory. *MRS Bulletin*, 31:681–687, 2006.

[8] G. Ceder. Predicting properties from scratch. *Science*, 280:1099, 1998.

[9] A. Van der Ven, M.K. Aydinol, G. Ceder, G. Kresse, and J. Hafner. First-principles investigation of phase stability in $li_xcoo_2$. *Physical Review B*, 58(6):2975–2987, 1998.

[10] G. Ceder, Y.M. Chiang, D.R. Sadoway, Aydinol M.K., Y.I. Jang, and B. Huang. Identification of cathode materials for lithium batteries guided by first-principles calculations. *Nature*, 392(6677):694–696, 1998.

[11] F. Zhou, M. Cococcioni, C. A. Marianetti, D. Morgan, and G. Ceder. First-principles prediction of redox potentials in transition-metal compounds with lda+u. *Physical Review B*, 70(235121), 2004.

[12] K. Kang, Y.S. Meng, J. Breger, C.P. Grey, and G. Ceder. Electrodes with high power and high capacity for rechargeable lithium batteries. *Science*, 311:977–980, 2006.

[13] M. Khantha, N.A. Cordero, L.M. Molina, J.A. Alonso, and L.A. Girifalco. Interaction of lithium with graphene: an *ab initio* study. *Physical Review B*, 70(125422), 2004.

[14] T. Mueller and G. Ceder. A density functional theory study of hydrogen adsorption in mof-5. *J. Phys. Chem. B*, 109:17974–17983, 2005.

[15] T. Mueller and G. Ceder. Effective interactions between the n-h bond orientations in lithium imide and a proposed ground-state structure. *Physical Review B*, 74(134104):1–7, 2006.

[16] C. Wolverton, V. Ozolins, and M. Asta. Hydrogen in aluminum: First-principles calculations of structure and thermodynamics. *Physical Review B*, 69(144109), 2004.

[17] C.R. Miranda and G. Ceder. Ab initio investigation of ammonia-borane complexes for hydrogen storage. *J. Chemical Physics*, 126(184703):1–11, 2007.

[18] V. Ozolins, E.H. Majzoub, and T.J. Udovic. Electronic structure and rietveld refinement parameters of ti-doped alanates. *J. Alloys and Compounds*, 375:1–10, 2004.

[19] B.C. Han and G. Ceder. Effect of coadsorption and ru alloying on the adsorption of co on pt. *Physical Review B*, 74(205418), 2006.

[20] G.H. Jóhannesson, T. Bligaard, A.V. Ruban, H.L Skriver, K.W. Jacobsen, and J.K. Nørskov. Combined electronic structure and evolutionary search approach to materials design. *Physical Review Letters*, 88(25), 2002.

[21] Y.-S. Lee and N. Marzari. Cycloaddition functionalizations to preserve or control the conductance of carbon nanotubes. *Phys. Rev. Lett.*, 116801, 2006.

[22] E.J. Wu and G. Ceder. Computational investigation of dielectric absorption at microwave frequencies in binary oxides. *J. App. Physics*, 89(10):5630–5636, 2001.

[23] J. Maddox. Crystals from first principles. *Nature*, 335:201, 1988.

[24] J. F. Nye. *Physical Properties of Crystals.* Oxford, 2 edition, 1985.

[25] M. Lazzeri and F. Mauri. Nonadiabatic kohn anomaly in a doped graphene monolayer. *Physical Review Letters*, 97(266407), 2006.

[26] R. M. Dreizler and E.K.U. Gross. *Density Functional Theory.* Springer-Verlag, Berlin, 1990.

[27] Nicola Marzari. *Ab-initio Molecular Dynamics for Metallic Systems.* PhD thesis, Pembroke College, University of Cambridge, 1996.

[28] M.C. Payne, M.P. Teter, D.C. Allan, T.A. Arias, and J.D. Joannopoulos. Iterative minimization techniques for *ab initio* total-energy calculations: molecular dynamics and conjugate gradients. *Reviews of Modern Physics*, 64(4):1045–1097, 1992.

[29] R.M. Martin. *Electronic Structure: basic theory and practical methods.* Oxford University Press, 2005.

[30] F. Aryasetiawan and O. Gunnarsson. The *GW* method. *Reports on Progress in Physics*, 61(3):237–312, 1998.

[31] G. Kotliar, S. Y. Savrasov, K. Haule, V. S. Oudovenko, O. Percollet, and C. A. Marianetti. Electronic structure calculations with dynamical mean-field theory. *Reviews of Modern Physics*, 78:865–951, 2006.

[32] D. Chandler. *Introduction to Modern Statistical Mechanics.* Oxford University Press, New York, 1987.

[33] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications.* Academic Press, 2 edition, 1996.

[34] K. Binder and D.W. Heermann. *Monte Carlo Simulation in Statistical Physics: An Introduction.* Series in Solid-State Sciences. Springer-Verlag, 1997.

[35] F. Ducastelle. *Order and Phase Stability in Alloys*, volume 3 of *Cohesion and Structure.* North-Holland, Amsterdam, 1993.

[36] D de Fontaine. Configurational thermodynamics of solid solutions. In H. Ehrenreich, F. Seitz, and D Turnbull, editors, *Solid State Physics: Advances in Research and Applications*, volume 34, pages 74–272. Academic Press, 1979.

[37] D. de Fontaine. Cluster approach to order-disorder transformations in alloys. In *Solid State Physics: Advances in Research and Applications*, volume 47, pages 33–176. Academic Press, 1994.

[38] W.H. Press. *Numerical Recipes in C: the art of scientific computing.* Cambridge University Press, 1992.

[39] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. Optimizing by simulated annealing. *Science*, 220:671, 1983.

[40] J.H. Holland. *Adaption in Natural and Artificial Systems.* big boy publishers, 1975.

[41] N.L. Abraham and M.I.J. Probert. A periodic genetic algorithm with real-space representation for crystal structure and polymorph prediction. *Physical Review B*, 73(224104), 2006.

[42] D.M. Deaven and K.M. Ho. Molecular-geometry optimization with a genetic algorithm. *Physical Review Letters*, 75:288–291, 1995.

[43] T.S. Bush, C.R.A. Catlow, and P.D. Battle. Evolutionary programming techniques for predicting inorganic crystal structures. *Journal of Materials Chemistry*, 5(8):1269–1272, 1995.

[44] C.W. Glass, A.R. Oganov, and N. Hansen. **uspex**-evolutionary crystal structure prediction. *Computer Physics Communications*, 175:713–720, 2006.

[45] G. Trimarchi and A. Zunger. Global space-group optimization problem: Finding the stablest crystal structure without constraints. *Physical Review B*, 75(104113), 2007.

[46] S.M. Allen and J.W. Cahn. Ground state structures in ordered binary alloys with second neighbor interactions. *Acta Metallurgica*, 20:423–433, 1972.

[47] J. Kanamori. Magnetization process in an ising spin system. *Progress in Theoretical Physics*, 35:66, 1966.

[48] G.D. Garbulsky. *Ground-state structures and vibrational free energy in first-principles models of substitutional-alloy thermodynamics.* PhD thesis, Massachusetts Institute of Technology, 1996.

[49] J. M. Sanchez, F. Ducastelle, and D. Gratias. Generalized cluster description of multicomponent systems. *Physica A*, 128:334–350, 1984.

[50] G. Ceder. A derivation of the ising model for the computation of phase diagrams. *Computational Materials Science*, 1:144–150, 1993.

[51] G. Ceder, G.D. Garbulsky, D. Avis, and K. Fukuda. Ground states of a ternary fcc lattice model with nearest and next-nearest-neighbor interactions. *Physical Review B*, 49(1):1–8, 1995.

[52] J.W.D. Connolly and A.R. Williams. Density-functional theory applied to phase transformations in transition metal alloys. *Physical Review B*, 27(8):5169–5172, 1983.

[53] G.D. Garbulsky and G. Ceder. Linear-programming method for obtaining effective cluster interactions in alloys from total-energy calculations: Application to the fcc pd-v system. *Physical Review B*, 51(1):67–72, 1995.

[54] V. Blum and A. Zunger. Prediction of ordered structues in the bcc binary systems of mo, nb, ta, and w from first-principles search of approximately 3,000,000 possible configurations. *Physical Review B*, 72, 2005.

[55] Z. Chen, Z. Lu, and J.R. Dahn. Staging phase transitions in $li_xcoo_2$. *Journal of the Electrochemical Society*, 149(12):A1604–A1609, 2002. in /home/ccfisch/NEED_TO_FILE.

[56] Y. Shao-Horn, S. Levasseur, F. Weill, and C. Delmas. *J. Electrochemical Society*, 150:A366, 2003.

[57] F. Zhou, T. Maxisch, and G. Ceder. Configurational electronic entropy and the phase diagram of mixed-valence oxides: The case of $li_xfepo_4$. *Physical Review Letters*, 97(155704), 2006.

[58] P.D. Tepesch, G.D. Garbulsky, and G. Ceder. Model for configurational thermodynamics in ionic systems. *Physical Review Letters*, 74(12):2272–2275, 1995.

[59] F. Zhou, G. Grigoryan, S.R. Lustig, A.E. Keating, G. Ceder, and D. Morgan. Coarse-graining protien energetics in sequence variables. *Physical Review Letters*, 95(148103):1–4, 2005.

[60] L. Pauling. The principles determining the structure of complex ionic crystals. *J. Am. Chem. Soc.*, 51:1010–1026, 1929.

[61] L. Pauling. *The Nature of the Chemical Bond.* Cornell University Press, third edition, 1960.

[62] F. Laves and H. Witte. unknown. *Naturwissen schaften*, 27:65, 1935.

[63] F.C. Frank and J.S. Kasper. Complex alloy structures regarded as sphere packings. i. definitions and basic principles. *Acta Cryst.*, 11:184, 1958.

[64] J.L.C. Daams, P. Villars, and J.H.N. van Vucht. Atomic-environment classification of the cubic intermetallic structure types. *J. Alloys and Compounds*, 182:1–33, 1992.

[65] W. Hume-Rothery. *J. Inst. Metals*, 35:307, 1926.

[66] W. Hume-Rothery. *Electrons, atoms, metals and alloys.* Dover Publications, 3 edition, 1963.

[67] N.F. Mott and H. Jones. *The Theory of the Properties of Metals and Alloys.* Clarendon Press, 1936.

[68] H. Jones. The phase boundaries in binary alloys, part 2: The theory of the $\alpha$, $\beta$ phase boundaries. *Proc. Phys. Soc.*, 49(3):243, 1937.

[69] A.T. Paxton, M. Methfessel, and D. Pettifor. A bandstructure view of the hume-rothery electron phases. *Proc. Roy. Soc. A*, 453(1962):1493, 1997.

[70] A.R. Miedema. *Philips Tech. Rev.*, 36:217, 1976.

[71] F.R. de Boer, R. Boom, W.C.M. Matten, A.R. Miedema, and A.K. Niessen. *Cohesion in metals: Transition Metal Alloys.* North Holland, 1988.

[72] E. Mooser and W.B. Pearson. On the crystal chemistry of normal valence compounds. *Acta Cryst.*, 12:1015–1022, 1959.

[73] W.A. Harrison. *Pseudopotentials in the Theory of Metals.* W.A. Benjamin, 1966.

[74] V. Heine and D. Weaire. Structure of di- and trivalent metals. *Phys. Rev.*, 152(2):603, 1966.

[75] G. Simons and A.N. Bloch. Pauli-force model potential for solids. *Phys. Rev. B*, 7(6):2754, 1973.

[76] Judith St. John and A. N. Bloch. Quantum-defect electronegativity scale for nontransition elements. *Phys. Rev. Lett.*, 33(18):1095, 1974.

[77] L. Brewer. Bonding and structure of transition metals. *Science*, 161(3837):115–122, 1968.

[78] A. Zunger. Systemization of the stable crystal structure of all ab-type binary compounds: A pseudopotential orbital-radii approach. *Phys. Rev. B*, 22(12):5839–5872, 1980.

[79] J.C. Phillips and J.A. Van Vechten. Dielectric classification of crystal structures, ionization potentials, and band structures. *Physical Review Letters*, 22(14):705–708, 1969.

[80] P. Villars. A three-dimensional structural stability diagram for 998 binary ab intermetallic compounds. *Journal of the Less Common Metals*, 92:215, 1983.

[81] P. Villars. *J. Less-Common Metals*, 119:175, 1986.

[82] P. Villars. Factors governing crystal structures. In J.H. Westbrook and R.L. Fleisher, editors, *Intermetallic Compounds: Vol. 1 Principles and Practice*, pages 227–275. John Wiley & Sons, 1994.

[83] O. Muller and R. Roy. *The Major Ternary Structural Families.* Crystal Chemistry. Springer, 1974.

[84] E.S. Machlin, T.P. Chow, and J.C. Phillips. Structural stability of suboctet simple binary compounds. *Physical Review Letters*, 38(22):1292–1295, 1977.

[85] J.K. Burdett and J.R. Rodgers. Structure & property maps for inorganic solids. In *Encyclopedia of Inorganic Chemistry.* Wiley, 2007.

[86] D. Morgan, J. Rodgers, and G. Ceder. Automatic construction, implementation, and assessment of pettifor maps. *J. Phys. C*, 15(25):4361, 2003.

[87] P. Villars, M. Berndt, K. Brandenburg, K. Cenzual, J. Daams, F. Hulliger, T. Massalski, H. Okamoto, K. Osaki, A. Prince, H. Putz, and S. Iwata. *PAULING FILE: Binaries Edition.* ASM International, Materials Park, Ohio, 2002.

[88] K.J. Tibbetts. Data mining for structure type prediction. Master's thesis, Massachusetts Institute of Technology, September, 2004.

[89] X.G. Gong, G.L. Chiarotti, M. Parinello, and E. Tosatti. $\alpha$-gallium: A metallic molecular crystal. *Physical Review B*, 43(17):14227–14280, 1991.

[90] K. Cenzual, J.L. Jorda, and E. Parthe. $Zr_3rh_5$ with $pu_3pd_5$-type structure, a structure geometrically related to the cscl type. *Acta Crystallographica Section C*, 44:14–18, 1988.

[91] T. Cover and J. Thomas. *Elements of Information Theory.* Wiley, New York, 1991.

[92] B.L. Henke, E.M. Gullikson, and J.C. Davis. X-ray interactions: photoadsorption, scattering, transmission, and reflection at $e = 50 \rightarrow 30000$ ev, $z = 1 \rightarrow 92$. In *Atomic Data and Nuclear Data Tables*, volume 54, pages 181–342. 1993.

[93] P. Coppens. *The structure factor*, volume B, chapter 1.2, pages 10–24. International Union of Crystallography, 2006.

[94] S.L. Lauritzen. *Graphical Models.* Oxford University Press, 1996.

[95] D. Koller and N. Freidman. *Bayesian Networks and beyond.* MIT CopyTech, forthcoming.

[96] David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 7 edition, 2003.

[97] M.J. Wainwright. *Stochastic processes on graphs with cycles: geometric and variational approaches.* PhD thesis, Massachusetts Institute of Technology, June 2002.

[98] M.I. Jordan, editor. *Learning in Graphical Models.* Adaptive Computation and Machine Learning. MIT Press, 1998.

[99] D. Heckerman. A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft Corporation, Redmond, WA, 1995.

[100] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, 1988.

[101] T. Morita. Cluster variation method of cooperative phenomena and its generalization i. *J. Physical Soc. Japan*, 12(7):753–755, 1957.

[102] R. Kikuchi. A theory of cooperative phenomena. *Physical Review*, 81(6):988–1003, 1951.

[103] J.G. Kirkwood and E.M. Boggs. The radial distribution function in liquids. *Journal of Chemical Physics*, 10:394–402, 1942.

[104] T. Morita. Formal structure of the cluster variation method. *Prog. of Theoretical Physics Supp.*, (115):27–39, 1994.

[105] G. An. A note on the cluster variation method. *J. Statistical Physics*, 52:727–734, 1988.

[106] A. G. Schlijper. Convergence of the cluster-variation method in the thermodynamic limit. *Physical Review B*, 27(11):6841–6848, 1983.

[107] W.J. McGill. Multivariate information transmission. *Psychometrika*, 19:97–116, 1954.

[108] M.V. Prokoféf, V.E. Kolesnichenko, and V.V. Karonik. Composition and structure of alloys in the mg-ag system near $mg_3ag$. *Inorganic Materials*, 21:1168, 1985.

[109] J. Kulik, S. Takeda, and D. de Fontaine. Long period superstructures in $ag_3mg$. *Acta Metall.*, 35(5):1137, 1987.

[110] O. Loebich and C.J. Raub. *J. Less-Common Metals*, 70(1):47–55, 1980.

[111] O. Loebich and C.J. Raub. Reactions between some alkali and platinum groum metals. *Platinum Metals Rev.*, 25(3):113–120, 1981.

[112] C.P. Nash, F.M. Boyden, and L.D. Whittig. Intermetallic compounds of alkali metals with platinum. a novel preparation fo a colloidal platinum hydrogenation catalyst. *J. American Chemical Society*, 82(6203-6204), 1960.

[113] W. Bronger, B. Nacken, and K. Ploog. Zur synthese und struktur von $li_2pt$ and lipt. *J. of the Less-Common Metals*, 43:143–146, 1975.

[114] W. Bronger, G. Klessen, and P. Mueller. Zur struktur von $lipt_7$. *J. of the Less-Common Metals*, 109:1–2, 1985.

[115] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Stat. Soc. B*, 36(2):111–147, 1974.

[116] J.A. Simmons. On the superposition of probabilities. In J.L. Moran-Lopez and J.M. Sanchez, editors, *Theory and Applications of the Cluster Variation and Path Probability Methods*, page 387. Plenum Press, 1996.

[117] M.J. Wainwright and M.I. Jordan. A variational principle for graphical models. In S. Haykin, J. Principe, T. Sejnowski, and J. McWhirter, editors, *New Directions in Statistical Signal Processing*, chapter 11. MIT Press, 2005.

[118] M.J. Wainwright and M.I. Jordan. Variational inference in graphical models: The view from the marginal polytope. In *Allerton Conference on Control, Communication and Computing*, October 2003.

[119] M.D. Asta. *First-Principles Calculations of Thermodynamic Properties and Phase Diagrams of Binary Substitutional Alloys.* PhD thesis, University of California at Berkeley, 1993.

[120] C.E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:623, 1948.

[121] E.T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

[122] E.T. Jaynes. *Probability Theory: the logic of science.* Cambridge University Press, 2003.

[123] R.T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, 1970.

[124] J.N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43:1470–1480, 1972.

[125] R. Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Conference on Natural Language Learning*, pages 49–55. Association for Computational Linguistics, 2002.

[126] R. M. Neal. Probabilistic inference using markov chain monte carlo methods. Technical report, Department of Computer Science, University of Toronto, 1993. http://www.cs.toronto.edu/ radford/.

[127] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Generalized belief propagation. In *Advances in Neural Information Processing Systems*, volume 13, pages 689–695. MIT Press, 2001.

[128] M.J. Wainwright, T. Jaakkola, and A.S. Willsky. Tree-based reparameterization framework for the analysis of sum-product and related algorithms. *IEEE Trans. on Information Theory*, 45(9):1120–1146, 2001.

[129] K. Murphy, Y. Weiss, and M.I. Jordan. Loopy belief propagation for approximate inference: An empirical study. volume 15 of *Proceedings of Uncertainty in AI*, 1999.

[130] M.J. Wainwright. Estimating the "wrong" graphical model: Benefits in the computation-limited setting. *Journal of Machine Learning Research*, 7:1829–1859, 2006.

[131] D.J. Spiegelhalter and R.P. Knill-Jones. Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *J. Royal Statistical Society. Series A*, 147(1):35–77, 1984.

[132] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[133] J. Binder, D. Koller, S. Russell, and K. Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29(2):213–244, 1997.

[134] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the *em* algorithm. *J. Roy. Stat. Soc. B*, 39(1):1–38, 1977.

[135] M.A.T. Figueiredo and A.K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–394, 2002.

[136] Peter Cheeseman and John Stutz. Bayesian classification (autoclass): Theory and results. In *Advances in Knowledge Discovery and Data Mining*, pages 61–83. AAAI Press, Menlo Park, 1996.

[137] G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6:461–464, 1978.

[138] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* Singapore World Scientific, 1989.

[139] J. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference*, Uncertainty in Artificial Intelligence, pages 43–52. Morgan Kaufmann, 1998.

[140] Greg Hamerly and Charles Elkan. Bayesian approaches to failure prediction for disk drives. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 202–209, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[141] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery, and W.L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(102001):977–987, 2001.

[142] R. Hundt, J.C. Schön, and M. Jansen. Cmpz – an algorithm for the efficient comparison of periodic structures. *J. App. Cryst.*, 39:6–16, 2006.

[143] G. Hautier. Affine mapping prototyping of the icsd. unpublished results, 2007.

[144] R. Cox. Probability, frequency, and reasonable expectation. *Am. J. Physics*, 14:1–13, 1946.

[145] T. Jaakkola. Machine learning lectures: 22. Technical report, MIT, 2006.

[146] W.M.C. Foulkes, L. Mitas, R.J. Needs, and G. Rajagopal. Quantum monte carlo simulations of solids. *Reviews of Modern Physics*, 73:33–83, 2001.

[147] L. Hedin. New method for calculating the one-particle green's function with application to the electron-gas problem. *Physical Review*, 139(3A):A796, 1965.

[148] G. Kresse and J. Furthmuller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science*, 6:15, 1996.

[149] Perdew. J. and Y. Wang. In P. Ziesche and H. Eschrig, editors, *Electronic Structure of Solids '91.* Akademie Verlag, 1991.

[150] J.P Perdew, K. Burke, and M. Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77(8):3865–3868, 1996.

[151] P.E. Blöchl. Projector augmented-wave method. *Physical Review B*, 50(17953), 1994.

[152] G. Kresse and J. Joubert. From ultrasoft pseudopotentials to the projector augmented wave method. *Physical Review B*, 59(1758), 1999.

[153] A.M. Rappe, K.M. Rabe, E. Kaxiras, and J.D. Joannopoulos. Optimized pseudopotentials. *Physical Review B*, 41(2):1227–1230, 1990.

[154] H.J. Monkhorst and J.D. Pack. Special points for brillouin-zone integrations. *Physical Review B*, 13:5188, 1976.