A materials informatics approach for crystal chemistry

Chang Sun Kong
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Materials Science and Engineering Commons

Recommended Citation
Kong, Chang Sun, "A materials informatics approach for crystal chemistry" (2009). Graduate Theses and Dissertations. 10831. https://lib.dr.iastate.edu/etd/10831

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
A materials informatics approach for crystal chemistry

by

Kong, Chang Sun

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Materials Science and Engineering

Program of Study Committee:
Krishna Rajan, Major Professor
Scott Chumbley
Zhiqun Lin
Derrick Rollins
James Oliver

Iowa State University

Ames, Iowa

2009

Copyright © Chang Sun Kong, 2009. All rights reserved.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TABLE OF CONTENTS</td>
<td>ii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>GLOSSARY</td>
<td>vii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>viii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>ix</td>
</tr>
<tr>
<td>CHAPTER 1</td>
<td></td>
</tr>
<tr>
<td>INTRODUCTION: DATA-DRIVEN APPROACHES IN MATERIALS RESEARCH</td>
<td>1</td>
</tr>
<tr>
<td>1.1. Classification in materials science</td>
<td>1</td>
</tr>
<tr>
<td>1.2. The structure of this thesis</td>
<td>8</td>
</tr>
<tr>
<td>References</td>
<td>12</td>
</tr>
<tr>
<td>CHAPTER 2</td>
<td></td>
</tr>
<tr>
<td>CRYSTALLOGRAPHIC DATA BASE: GROUP-THEORETICAL CLASSIFICATION OF AB₂-TYPE STRUCTURES</td>
<td>14</td>
</tr>
<tr>
<td>2.1. Linus Pauling File data</td>
<td>14</td>
</tr>
<tr>
<td>2.2. Hierarchical crystal-symmetry classification</td>
<td>15</td>
</tr>
<tr>
<td>References</td>
<td>31</td>
</tr>
<tr>
<td>CHAPTER 3</td>
<td></td>
</tr>
<tr>
<td>DESIGN RULES OF CRYSTAL STRUCTURES USING RECURSIVE PARTITIONING METHODS</td>
<td>32</td>
</tr>
<tr>
<td>3.1. Introduction</td>
<td>32</td>
</tr>
<tr>
<td>3.1.1. The mapping of materials information</td>
<td>32</td>
</tr>
<tr>
<td>3.1.2. Recursive partitioning and classification tree</td>
<td>39</td>
</tr>
<tr>
<td>3.2. Methods</td>
<td>43</td>
</tr>
<tr>
<td>3.2.1. Construction of data set</td>
<td>43</td>
</tr>
<tr>
<td>3.2.2. Selection of atomic parameters</td>
<td>43</td>
</tr>
<tr>
<td>3.2.3. Information entropy</td>
<td>45</td>
</tr>
<tr>
<td>3.2.4. Classification tree: Partitioning of multidimensional space</td>
<td>46</td>
</tr>
<tr>
<td>3.3. Results and discussion</td>
<td>49</td>
</tr>
<tr>
<td>References</td>
<td>60</td>
</tr>
</tbody>
</table>
CHAPTER 4
INFORMATION-ENTROPY SCALED STRUCTURE MAPS ........................................ 64
  4.1. Introduction ............................................................................................................ 64
  4.2. Computational details ........................................................................................... 70
    4.2.1. Classification of data .................................................................................... 70
    4.2.2. Classification by information entropy ....................................................... 71
  4.3. Results and discussion ........................................................................................... 75
    4.3.1. Assessment of atomic, physical parameters by information entropy ... 75
    4.3.2. Classification rule mining of crystal chemistry ........................................ 79
References....................................................................................................................... 87

CHAPTER 5
APPLICATION TO MATERIALS DESIGN...................................................................... 90
  5.1. Integration of recursive partitioning and first principles calculations:
      prediction of the stable crystal structures .......................................................... 90
    5.1.1. Energy calculation by density functional theory ..................................... 91
    5.1.2. Prediction of stable crystal structures ........................................................ 92
  5.2. Phase-transition paths of metal-hydride alloys ................................................. 94
References..................................................................................................................... 100

CHAPTER 6
CONCLUSIONS: CONTRIBUTIONS OF THIS STUDY .............................................. 101

APPENDIX I
MATHEMATICS OF CLASSIFICATION RULE MINING ........................................... 103
  A1.1. Introduction ....................................................................................................... 103
  A1.2. Classification tree .............................................................................................. 104
    A1.2.1. Solving for the best partitions: Impurity function .............................. 105
  A1.3. Algorithmic framework.................................................................................... 107
    A1.3.1. Recursive partitioning of a two-class problem................................. 107
    A1.3.2. Recursive partitioning of a multi-class problem ............................... 110
References..................................................................................................................... 114

APPENDIX II
CRYSTALLOGRAPHIC GROUP-SUBGROUP RELATIONS...................................... 115
  A2.1. Space group symbols ........................................................................................ 115
  A2.2. Symmetry reduction between space groups .................................................. 117
References..................................................................................................................... 119
LIST OF FIGURES

Figure 1.1. Pattern recognition of entities in the feature space........................................ 3
Figure 1.2. A schematic example of the reduction of uncertainty of a system by classification task................................................................................................................... 4
Figure 1.3. Mendeleev’s periodic table................................................................. 5
Figure 1.4. A schematic of the feature space of the structure-governing factors......... 6
Figure 1.5. Some examples of structure maps. ................................................................. 7
Figure 1.6. An example of the chemical bond triangle. ........................................ 8
Figure 1.7. Two approaches, crystal chemistry and crystallography, for the study of crystal structures and physical properties of materials systems. ........................................ 9
Figure 1.8. The structure of this thesis. ............................................................................. 11
Figure 2.1. Linus Pauling File (LPF) database................................................................. 15
Figure 2.2. The hierarchy scheme of the crystallographic symmetry............................. 17
Figure 2.3. Classification of the AB$_2$-type crystal structures based on the symmetrical hierarchy............................................................................................................................... 20
Figure 2.4. The relationships among crystal system, point group, and space group is shown................................................................................................................................. 22
Figure 2.5. The relationship of symmetry groups of a hexagonal system..................... 24
Figure 2.6. The symmetry relations between two structure prototypes, AlB$_2$ and CaIn$_2$, and the symmetry reduction ................................................................................................................................. 26
Figure 2.7. The relations between a (super) space group P6/mmm (for the structure types, AlB$_2$ and Hg$_2$U) and their maximal subgroups................................................................. 27
Figure 2.8. The relationship of symmetry groups of a cubic system............................. 28
Figure 2.9. The relations between a (super) space group P6/mmm (for the structure types, AlB$_2$ and Hg$_2$U) and its maximal t-subgroup I4/mmm (for the structure types, CaC$_2$, CuZr$_2$, LaSb, and MoSi$_2$) with its index equals 3................................................................. 28
Figure 2.10. The symmetry relations between structures are refined by organizing group-subgroup relationships from the two starting structure types (CaF$_2$ and AlB$_2$/Hg$_2$U) correspond to the cubic and hexagonal systems, respectively (compare with Figure 2.3)................................................................................................................................. 29
Figure 2.11. The symmetrical hierarchy of crystal-structure types of AB$_2$ system. .... 30
Figure 3.1. From the Periodic Table to multivariate structure map .................... 36
Figure 3.2. Scatter plot of seven physical parameters ................................. 39
Figure 3.3. (a) Information entropy change (ΔH) by the partitioning of the parameter space, and (b) a tree structure for the graphical representation of the classification result ................................................................. 40
Figure 3.4. Illustration of the partitioning of parameter space ...................... 42
Figure 3.6. One-dimensional structure mapping. ....................................... 45
Figure 3.7. A classification tree for the crystal structures of AB<sub>2</sub> compounds ... 53
Figure 3.8. Partitioning of chemical space that includes AlB<sub>2</sub> structure domain ... 56
Figure 4.1. A schematic of (a) data classification (i.e. partitioning) in the 2-D parameter space (X<sub>1</sub> and X<sub>2</sub>) and (b) its classification-tree expression ..................... 71
Figure 4.2. Evaluation of atomic and physical parameters from three structure-map models .................................................................................................................. 79
Figure 4.3. The contribution of atomic-property parameters to the respective structure types .............................................................................................................. 82
Figure 4.4. Information-entropy diagram: the relative contributions of structure-governing factors for AB<sub>2</sub> compounds ................................................................. 83
Figure 4.5. Information-entropy diagram .................................................. 85
Figure 5.1. The integration of two principles for the estimation of the stable crystal structure ........................................................................................................ 91
Figure 5.2. Prediction of the crystal structure of AuBe<sub>2</sub> .................................. 93
Figure 5.3. A schematic description of the structural changes according to the hydrogenation-dehydrogenation cycling process ........................................ 96
Figure 5.4. A part of the classification tree for AB<sub>2</sub> compounds ................. 98
Figure 5.5. The symmetrical relations between the crystal structure of Mg<sub>2</sub>FeH<sub>6</sub> and a hypothetical structure of the corresponding host Mg<sub>2</sub>Fe compound ................. 99
Figure A1.1. Recursive partitioning ........................................................... 104
Figure A1.2. Impurity functions ........................................................................ 106
Figure A1.3. Histograms for two different parameters (ΔX<sub>MB</sub> and ΔR<sub>s+p</sub>) for the CuAl<sub>2</sub> (tetragonal, I4/mcm) and KHg<sub>2</sub> (orthorhombic, Imma) structure types .......... 110
Figure A1.4. The partitioning of parameter space and the corresponding classification tree ........................................................................................................ 111
Figure A1.5. The schematic procedure of the data classification by using recursive partitioning for the application to the crystal chemistry research .................. 113
LIST OF TABLES

Table 2.1. The order of the point groups ................................................................. 18
Table 2.2. Crystal systems and the relevant point group symmetries .................... 21
Table 2.3. Crystallographic data of 34 crystal structure types of AB$_2$ systems ....... 23
Table 3.1. Atomic and physical parameters for the crystal structure prediction ....... 38
Table 3.2. If-then rules for the prediction of possible crystal structures .............. 57
Table 4.1. A brief and partial list of the references on the factors governing crystal structures .................................................................................................................. 65
Table 4.2. Atomic parameters for the classification of crystal structure data ......... 76
Table 4.3. The summary of the results of information-theoretic classification of AB$_2$ compounds ......................................................................................................................... 84
Table 5.1. Prediction of the crystal structure (prototype) of AB$_2$ type compounds ... 94
Table A1.1. Determination of the partitioning position by information-entropy calculations ................................................................................................................................. 109
Table A2.1. Crystallographic space groups ........................................................... 116
### GLOSSARY

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>APP</td>
<td>Atomic, physical parameter</td>
</tr>
<tr>
<td>argmax</td>
<td>Argument of the maximum</td>
</tr>
<tr>
<td>EC</td>
<td>Electrochemical factor</td>
</tr>
<tr>
<td>H</td>
<td>Information entropy</td>
</tr>
<tr>
<td>ΔH</td>
<td>The reduction of the entropy</td>
</tr>
<tr>
<td>IG</td>
<td>Information gain</td>
</tr>
<tr>
<td>m</td>
<td>A node of the classification tree</td>
</tr>
<tr>
<td>p(x)</td>
<td>The unconditional probability of x</td>
</tr>
<tr>
<td>p(y/x)</td>
<td>The conditional probability of y given x</td>
</tr>
<tr>
<td>p(x,y)</td>
<td>The joint probability of x and y</td>
</tr>
<tr>
<td>p_m(x)</td>
<td>The probability of the class $x_i$ in the node m</td>
</tr>
<tr>
<td>p_{mk}</td>
<td>The proportion of class k observations in a node m</td>
</tr>
<tr>
<td>Q_m(T)</td>
<td>The general form of the measures for the impurity of the node m of a tree T</td>
</tr>
<tr>
<td>SZ</td>
<td>Size factor</td>
</tr>
<tr>
<td>VE</td>
<td>Valence-electron number factor</td>
</tr>
<tr>
<td>x_i(m)</td>
<td>A major class (i.e., a structure type) in a node m</td>
</tr>
</tbody>
</table>
Thanks be to God. This thesis could not be completed without the support of many people and it is my great pleasure to thank those who made this thesis possible.

I would like to express all my gratitude to my advisor Professor Krishna Rajan. I am very thankful to him for his encouragement, guidance and all the supports for my research works. I am grateful to Dr. Changwon Suh. I am indebted to him for all his helpful discussion and advice for this thesis. I thank my friend Scott Broderick. Working and ALC (After Lunch Coffee break) with him have always been wonderful time to me. I also thank to all other my former and current group members for their help.

I owe my deepest gratitude to my wife, Seung Ha Kim. All my work would not have been possible without her help. I am the happiest man in the world since you are with me. We will be together all the time, my love. I am grateful to my brother Ick-Sun Kong. As Rocky says, “let’s keep going forward.” Never forget your dream, my brother!

This thesis is dedicated to my parents. All of this is as much as a product of their love and sacrifice for me as it is of mine. Just as you know me so well, I wish you know how much I love you. Thank you very much.
ABSTRACT

This thesis addresses one of the fundamental questions in materials crystal chemistry, namely why do atoms arrange themselves in the way they do? The ability to broadly design and predict new phases [i.e. crystal structures] can be partly met using concepts that employ phase homologies. Homologous series of compounds are those that seem chemically diverse but can be expressed in terms of a mathematical formula that is capable of producing each chemical member in that crystal structure. A well-established strategy to help discover new compounds – or at least to try to develop chemical design strategies for discovery – is to search, organize and classify homologous compounds from known data. These classification schemes are developed with the hope that they can provide sufficient insight to help us forecast with some certainty, specific new phases or compounds. Yet, while the classification schemes (over a dozen have been reported in the last 50 years) have proved to be instructive, mostly in hindsight, but they have had limited impact, if at all, on the a priori design of materials chemistry.

The aim of this research project is to develop a totally new approach to the study of chemical complexity in materials science using the tools of information theory and data science, which link diverse and high dimensional data derived from physical modeling and experiments. A very large scale binary $AB_2$ crystallographic database is used as a data platform to develop a new data mining / informatics protocol based on high dimensional recursive partitioning schemes coupled to information theoretic measures to:

- Identify which type of structure prototype is preferred over another for a given chemistry of compound
- discover new classification schemes of structure/chemistry/property relationships that classical homologies do not detect and finally we
- Extract and organize the underlying design rules for the formation of a given structure by quantitatively assessing the influence of multidimensional electronic structure attributes.

Finally some applications of this new approach are demonstrated; including new ways for linking first principles calculations to crystal structure prediction and group theory to crystal structure transition.
CHAPTER 1
INTRODUCTION: DATA-DRIVEN APPROACHES IN MATERIALS RESEARCH

1.1. Classification in materials science

Classification has long been an essential scientific tool to seek the underlying trends, rules, relationships, etc. of entities and systems, aiding the comprehensive understanding of the relevant complexities therein. It is on this framework of classification that materials science and research were founded. Any material can be classified into metal, semiconductor, or insulator, based on its electrical property. In the particle physics, a newly-discovered particle can be categorized into one of the classes of particles such as fermions and bosons within the systematization based on the spin nature of particles. The bonding between atoms of the same or different kinds in the condensed phase can be characterized by three types of states, that is, ionic, covalent, and metallic states; in fact, the actual interaction between atoms has a unique state that can be identified as an intermediate propensity somewhere in between the extreme three bonding states. It is the main objective of this study that we extract the useful information and knowledge for the materials design from the classification of the existing experimental database. This chapter introduces the conception of the classification in the materials research using some well-known examples, that is, the Periodic Table of chemical elements, structure maps for multi-element systems such as intermetallic compounds, and triangular diagrams which present the nature of chemical interactions between atomic elements in materials.
The examples provided reflect the descriptive aspect of the classification scheme through, namely, the “mapping of materials information.” It is the underlying fundamental principles of materials in which the regularities represent in terms of such as the periodicities, patterns, and rules induced from the classification of data.

In classification, any pattern embedded in data can be recognized in a supervised or unsupervised way. To understand what is meant by “the recognition of pattern,” a general description is given here. The spatial arrangement of entities in a feature space reveals the similarity of those which are like by the relative proximity (see the first row of Figure 1.1). In this case, the entities closely positioned form a closed region that can be called as a “domain” and the same class membership is assigned for the entities involved in the same domain. If certain conditions or properties can be used as the criterion for assigning memberships, the classification-based modeling simply is to find the boundaries between different domains (the second row of Figure 1.1). In reality, the activity of classification – whether it is supervised or unsupervised – as finding the boundaries for the separation of dissimilar classes is not trivial. Selection of the appropriate set of attributes forming feature space is also an important task as a part of data classification. A variety of data mining techniques provide the criteria of classifiers in the multi-dimensional feature space. Once the boundaries of different domains are determined i.e. the class membership of each entity in the data is given, the classification model can be used for 1) the prediction of the membership of new entities which were not employed for building the model, and for 2) the correlation of the entities with the attributes which specify a certain feature space.
Figure 1.1. Pattern recognition of entities in the feature space; the patterns – the grouping/separation of like or unlike entities – of spatial arrangements of entities shown here have one of the three types, i.e. (a) regular, (b) random, and (c) clustered. Whereas from the three cases on the first row the pattern of the entities is uncovered by considering the spatial arrangement of the entities, in the second row it is recognized by both the spatial arrangement and other conditions or properties (e.g. those denoted by red and blue colors) given a priori. In both cases of first and second row, the determination of dissimilar domains, i.e. classes or clusters, is a critical component of the pattern recognition.

Figure 1.2 describes the concept of uncertainty function (in other words, impurity, entropy function, etc.) as the qualitative and quantitative measure of the extent of the classification of a system (i.e. data). When the data (blue or red dots) are completely sorted out according to their similarity (the color), the uncertainty is decreased (since, for example in this Figure 1.2, the probability that one selects a dot with the color he wants would be increased from 50 to 100% !). The details of this conception, especially on the information entropy, will be dealt later in this thesis.
Figure 1.2. A schematic example of the reduction of uncertainty of a system by classification task. As the numerical measure of the uncertainty, a continuous function of the probability ($p_i$, where, $i =$ red or blue) can be defined; the uncertainty is minimized when the red and blue dots are completely separated (the box left side) and it is maximized when they are completely mixed together (the box right side).

The first example of the mapping of materials information is the Periodic Table of chemical elements. Dmitri Mendeleev, since 1869, has made a series of remarkable prediction of the existence of as-yet-unknown elements in his day, e.g. scandium, gallium, and germanium, using his systematic tabulation of the data of then-known elements. From the regularities found out in the table, namely periodic nature of the chemical elements, he was also able to propose the estimation of the physical/chemical properties of the predicted elements that later turned out to be closely matched with the experimental results. Considering the fact that his prediction was done prior to the advent of modern atomic model, the capability Periodic Table showed must have inspired all the materials researchers to date as well as Mendeleev’s contemporaries; it might be quite natural that the success of the Periodic Table for elements has provoked us to search such type of classification

† The existence of electron, proton and neutron was discovered in 1897, 1911, and 1932, respectively.
scheme, in general referred to as a “structure map”, for the multi-elements systems (i.e. multinary alloys, compounds, etc.) [1.1].

![Mendeleev's periodic table](image)

**Figure 1.3.** Mendeleev’s periodic table. Mendeleev published his first periodic table in 1869, in which the systematic collection of the information of chemical elements known in his age has been laid out in a table.

Structure map is a diagram that the regularities in the formability or structural stability of compounds are graphically represented by the correlation of the existing structure data with appropriate physical, or phenomenological, factors. When using the factors as the coordinates of the plot, the respective compounds are expected to be positioned - each compound is shown as a point in this plot - in a way that the similarity between the compounds is represented by the vicinity (i.e., the distance) among them. As the result, the compounds with a same structure fall into a certain range of each physical coordinate. In actual structure maps, there often exist some overlapping and/or outlying examples which hinder the clear determination of the borders between different structure regions. Previous structure-mapping approach can be categorized into two types according to the kind of the parameters used for the coordinates of structure map. One class of maps is to use two or three physical parameters, e.g. the ratio of atomic sizes or the difference of electronegativities of the constituent elements, as the orthogonal coordinates which would effectively separate out the compounds according to their crystal structures. Indeed, these
physical-coordinate maps have successfully classified some of particular compound groups within a two- or three-dimensional layout [1.1-1.6]. However, it has been recognized that the consideration of more than four factors is necessary for the determination of the structural stability of compounds [1.7]. It means that at least a four-coordinates plot should be built for the multi-elements systems, in which the graphical visualization and interpretation become more complicated. Instead, other methods which use phenomenological parameters designed for the best separation of different structure types were proposed. By assigning one or two phenomenological indices for each element, the compound information can be rendered on a two-dimensional plot. Chemical scale and Mendeleev number introduced by D. G. Pettifor [1.8] and elemental-property parameters derived from atomic number and periodic number by P. Villars [1.9] are the example.

**Figure 1.4.** A schematic of the feature space of the structure-governing factors. The multidimensional nature of data requires the multivariate analysis/modeling algorisms and methods.
Figure 1.5. Some examples of structure maps. Structure-mapping models suggested by (a) Mooser-Pearson (1959), (b) Miedema (1973), (c) Villars (1983), and Makino (1994). Most of previous structure maps were built with one-, two-, and three-parameter coordinates.

The interactions of atoms in the condensed phase are characterized by some bonding types, i.e. ionic, metallic, and covalent bonds. The following table shows the possible materials which can be achieved by the combination of first-row elements in the Periodic Table. As the list of materials is combined with the regularities embedded in the Periodic Table, one can find the characteristic relationships between the periodicity of chemical elements and the interaction (bonding) of atomic constituents in the materials.
Figure 1.6. An example of the chemical bond triangle [1.10]. Combinatorial table is constructed by the combination of the main group elements of the first row of the periodic table. The position of each compound denotes the bonding type of the compound.

1.2. The structure of this thesis

In this section, a short overview of this thesis is given along with the description of the interconnection between them. As shown schematically in the Figure 1.7, the foundations of this study of the crystal structure of condensed phases are sought from the crystal chemistry and crystallography using data-driven approaches as the
effective tool.

Chapter 1 began with some examples as to the data classification applied for materials research. Periodic Table was described as a classification/mapping of the information of chemical elements. As the extension of the Periodic Table to the multi-elements systems, it was pointed out that how structure maps have been developed and what is their limitation. In addition, multivariate data classification was emphasized as a next version of structure map.

Chapter 2 describes an interpretation of crystal structure data by crystallographic group theory. The similarity and connection between different crystal structures are examined in terms of symmetry relations, and they are applied as a virtual screening criterion for novel materials development.

![Diagram](image)

**Figure 1.7.** Two approaches, crystal chemistry and crystallography, for the study of crystal structures and physical properties of materials systems.

Chapter 3 is dedicated to the application of the classification of multivariate data to
crystal structure prediction. It is shown that sets of rules derived from the classification of the crystal structure data are used as the guideline for the quantum mechanical calculations.

Chapter 4 deals with classification rule mining by information theory. Crystal-chemistry rules are quantitatively assessed by information entropy measure during the classification of data.

Chapter 5 explores the aspects of application of the results that have shown at the previous chapters. The design rules for predicting the stable crystal structure of new compounds and the structural guideline of hydrogen storage materials are sought from the classification tree and structure family tree built from the materials data.

Chapter 6 concludes the thesis with the summary of the contributions and importance of this study on the materials research. The following diagram shows the structure of this thesis.

In this thesis, the regularities embedded in the structure formation of crystalline solids are investigated by means of a materials informatics approach, the multivariate classification of the crystallographic database of inorganic compounds. The regularities uncovered – i.e. classification rules extracted from materials data – are developed as the guideline for novel materials design.
Figure 1.8. The structure of this thesis.
References


CHAPTER 2

CRYSTALLOGRAPHIC DATA BASE: GROUP-THEORETICAL
CLASSIFICATION OF AB$_2$-TYPE STRUCTURES

2.1. Linus Pauling File data

The Linus Pauling File (LPF) is a huge volume of materials database constructed through an international collaborative project starting from 1995. The compilation includes about 200,000 entries of structure, property, diffraction, and constitution data of all kinds of inorganic solid phase materials (Figure 2.1). This comprehensive collection for the systematics of materials data allows scientists to readily search useful information by simple (and also complex) queries. However, the “search and match” type usage of this electronic repository of materials data does not generate any understanding or knowledge as to the materials behaviors – related to such as structures and properties – although that is the outcome of the long efforts is ultimately aiming at. In this thesis, the crystal structure data of binary compounds in the LPF database have been used to find the hidden regularities in diverse materials systems. This chapter deals with the crystallographic group theoretical classification of the crystal structures of AB$_2$ compounds, and data-mining methods will be applied from the next chapters.
Figure 2.1. Linus Pauling File (LPF) database [2.1]. This materials database includes approximately 80,000 structural data of multinary materials (28,300 for binaries). This figure was modified from the reference 2.1.

2.2. Hierarchical crystal-symmetry classification

The geometrical ordering of (chemical elemental) atoms in the solid state can be described systematically by the formalism of crystallography. This classification in terms of crystallography allows one to see diverse crystal structures of solids in terms of the similarity in the symmetry, geometrical dimension and relative positions. The prototype of crystal structure is therefore defined by the following conditions; the space group, same (or similar) ratio of the lattice parameters, Wyckoff positions. In addition, the correlation of different crystal structures by using
the symmetrical considerations of “(super) group-subgroup relations” provides the hierarchical connections of the crystal structures.

The relationships between different crystal structures are instrumental information for the design of novel materials. The connections – the similarity and hierarchy in a symmetry viewpoint – among different crystal structures can be effectively presented by crystallographic group-subgroup relations. The relations are compiled in a structure-family trees referred to as a Bärnighausen tree. In this study, the family tree as to the AB$_2$ binary compounds was constructed for the investigation of the structural similarities among the prototypes.

As the necessary data for building a structure-family tree, the symmetrical relations between space groups became available in full from the International Tables for Crystallography, Vol. A and A1 [2.2, 2.3]. The specific information for the investigation of group-subgroup relations, which can be obtained from these systematic tables, includes:

- Atomic coordinates
- Maximal klassengleiche/translationengleiche subgroups
- Minimal translationengleiche/non-isomorphic klassengleiche supergroups
- Transformation index
- Wyckoff positions

Figure 2.2 shows the crystallographic classification of the crystalline structures. Using this scheme for the description of crystal structure, the symmetries of any structure of a material not only can be determined in a systematic manner, but the symmetry relations among the crystal structures in three-dimensional space can also be examined effectively. At first, it will be shown that how the relations between two space-group types (defined by Hermann-Mauguin symbol) are determined starting
from this classification. Then, the group-subgroup relations of 34 structure types for AB₃ compounds will be derived in the form of a Bärnighausen tree along with the explanation on the procedure using some examples.

![Crystal symmetry hierarchy](image)

**Figure 2.2.** The hierarchy scheme of the crystallographic symmetry; the 230 space groups are based on the corresponding point groups along with the introduction of screw axes and glide planes.

Space groups G₁ and G₂ have symmetry relations of supergroup-subgroup if a set of symmetry operations of a space group (i.e. G₂) is the subset of those of another space group (i.e. G₁); in this case, G₁ is referred to as a “supergroup” and G₂ is referred to as a “subgroup.” Especially, if there are no subgroups between G₁ and G₂, the subgroup G₂ is the maximal subgroup of G₁. The symmetry relations – i.e. symmetry reduction from G₁ to G₂ (i.e., G₁ > G₂) – of these two space groups are presented using an arrow shown in the following expression:

\[ G₁ → kn \text{ or } tn \rightarrow G₂ \]

Where, G₁ is a supergroup, i.e. a space group with a higher symmetry and G₂ is a subgroup, i.e. a space group with a lower symmetry. Symbols k and t denote the
subgroups of the “same class” (=\textit{klassengleiche} subgroup) and those with the “same translations” (=\textit{translationengleiche} subgroup), respectively. From here, \(k\)-subgroup and \(t\)-subgroup will denote the \textit{klassengleiche} and \textit{translationengleiche} subgroup, respectively. Also, integer number \(n\) is the index which indicates the symmetry reduction between \(G_1\) and \(G_2\). This index is the number of cosets of subgroup \(G_2\) in the supergroup \(G_1\), implying the relative size of \(G_2\) on \(G_1\). When the space groups of two structures have group-subgroup relations, the index, \(n\) is calculated by the following equation (due to the Hermann theorem).

\[
 n = n_k \cdot n_t = \frac{M(G_2)}{M(G_1)} \cdot \frac{P(G_1)}{P(G_2)}
\]

(2.1)

Where, \(n\), \(n_k\), and \(n_t\) are the index of a transformation between \(G_1\) and \(G_2\) (\(G_1 > G_2\)), the \textit{klassengleiche} index, and \textit{translationengleiche} index, respectively. The index, \(n\) is a prime number for maximal subgroups, i.e. subgroups with the smallest step of symmetry reduction. \(M(G_1)\) and \(M(G_2)\) are the multiplication of primitive unit cell, i.e. the number of formula units per primitive unit cell, respectively. \(P(G_1)\) and \(P(G_2)\) are the orders of the point groups of the corresponding space group \(G_1\) and \(G_2\), respectively (Table 2.1):

\begin{table}[h]
\begin{tabular}{ccc}
1: & 1 & 6: 32, 3m \\
2: \(\bar{1}, 2, m\) & 8: mmm & 12: 23 \\
3: & 3 & 8: 4/m \\
4: \(2/m, 222, mm2\) & 8: 422, 4mm, \(\bar{4}2m\) & 24: \(m\bar{3}\) \\
4: & 4, \(\bar{4}\) & 12: 6/m \\
6: \(\bar{3}, 6, \bar{6}\) & 12: \(\bar{3}m, 622, 6mm, \bar{6}2m\) & 48: \(m\bar{3}m\) \\
\end{tabular}
\end{table}
Using this concept of the hierarchy of symmetry, symmetry reduction, and structural similarities based on group theory, the relations of 34 structure types of AB₂ compounds have been derived and presented graphically in a form of structure-family network.

As the first step of the investigation of symmetry relations, the prototypes of crystal structures of AB₂ compounds are classified according to the corresponding crystal system. The 34 structure types fall into the one of five crystal systems (cubic, hexagonal/trigonal, tetragonal, orthorhombic, and monoclinic) among the total seven systems (Figure 2.3); note that two space groups given do not have to be belonged to the same crystal system for achieving group-subgroup relations. Two hexagonal systems, AlB₂ and Hg₂U (P6/mmm, space-group number 191) and a cubic system CaF₂ (Fm̅3m, 225) have the highest symmetry and become the two starting points of the structure-family tree; the space groups, P6/mmm and Fm̅3m correspond to the highest symmetry group in hexagonal and cubic crystal system, respectively. Thus, Bärnighausen tree of crystal structures begins from the hexagonal system of AlB₂, Hg₂U and cubic system of CaF₂ as the aristotypes (i.e., basic structure), and are extended to the hettotype structures (i.e., derivative structures) which have the lower symmetry. Although a space group generally has multiple, or infinite, levels of lower symmetry space groups as the subgroups, only maximal subgroups are of interest.
Figure 2.3. Classification of the AB₂-type crystal structures based on the symmetrical hierarchy. The symmetry is reduced along the direction of arrows; for instance, cubic system has higher symmetry than monoclinic system.

To illustrate the procedure of generating group-subgroup relations, here the first step of the symmetry reduction from two aristotype space groups, P6/mmm (hexagonal system) and Fm̅3m (cubic system) will be presented. Consider eight structure types that belong to the hexagonal system. These eight structure types are subdivided into five different space groups; that is,

- AlB₂, Hg₂U (P6/mmm, 191),
- CaIn₂, MgZn₂, Ni₃In (P6₃/mmc, 194),
- Fe₃P (P6̅2m, 189),
- Cd₃Ce (P3̅m1, 164), and
- CrSi₂ (P6₃22, 180)
Some finite point group symmetries are allowed for each crystal system. The symmetry elements of the corresponding point groups are listed in increasing order in the Table 2.2; that is, the symmetry is increased from left to right.

Table 2.2. Crystal systems and the relevant point group symmetries.

<table>
<thead>
<tr>
<th>Crystal system</th>
<th>Point group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cubic</td>
<td>23, m$ar{3}$, 432, 4 3m, m$ar{3}$m</td>
</tr>
<tr>
<td>Hexagonal</td>
<td>6, 6, 6/m, 622, 6mm, 6 m2, 6/mmm</td>
</tr>
<tr>
<td>Trigonal*</td>
<td>3, 3, 32, 3m, 3m</td>
</tr>
<tr>
<td>Tetragonal</td>
<td>4, 4, 4/m, 422, 4mm, 4 m2, 4/mmm</td>
</tr>
<tr>
<td>Orthorhombic</td>
<td>222, mm2, mmm</td>
</tr>
<tr>
<td>Monoclinic</td>
<td>2, m, 2/m</td>
</tr>
</tbody>
</table>

* A subsystem of hexagonal system

Figure 2.4 presents that how crystal systems, crystal classes (point groups), and space groups are related each other. In hexagonal system, since the symmetry of point group is increased in the order of 6, 6, 6/m, 622, 6mm, 6 m2, and 6/mmm, the relative symmetry relation of space groups is determined by this order of symmetry (i.e. 6 → 6 → 6/m → 622 → 6mm → 6 m2 → 6/mmm) in the corresponding point groups. Among the space-group types belonging to the same crystal class (i.e. k-subgroups) have the hierarchical relationships. In fact, the hierarchy of the structure types belonging to the hexagonal system, which basically requires the systematic one-by-one search, is investigated readily by using the full listing of subgroups for each space group provided in the International Tables for Crystallography, Vol. A1 [2.3]. However, as U. Müller indicated [2.4], the group-subgroup relations purely formed with space-group information may not be so useful unless the exact atomic-
coordinates data are not known. Thus, for the practical application to real materials, more detailed examination on the atomic arrangement of each material system would be required. The information on the atomic positions of the respective structure types can be found from literatures, e.g., reference 4 on the web site. In the Table 2.3, the crystallographic information of AB₂ type structures are summarized.

![Figure 2.4](image_url)

**Figure 2.4.** The relationships among crystal system, point group, and space group is shown; for clarity, only the maximal k-subgroups are presented for the space group.
Table 2.3. Crystallographic data of 34 crystal structure types of AB$_2$ systems.

<table>
<thead>
<tr>
<th>No.</th>
<th>Structure type</th>
<th>Pearson symbol</th>
<th>Space group</th>
<th>Space group number</th>
<th>The number of unit cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AlB$_2$</td>
<td>hP3</td>
<td>P6/mmm</td>
<td>191</td>
<td>45</td>
</tr>
<tr>
<td>2</td>
<td>CuCo</td>
<td>hP6</td>
<td>P4/mnm</td>
<td>133</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>CeF$_2$</td>
<td>cP12</td>
<td>Pm-3m</td>
<td>224</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>Ceh$_2$</td>
<td>hP6</td>
<td>P6$_3$/mmc</td>
<td>194</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>Ce$_2$Co</td>
<td>hP3</td>
<td>P-3mart</td>
<td>184</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>Co$_2$P</td>
<td>cP12</td>
<td>Prwma</td>
<td>62</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>Co$_2$Sb-b</td>
<td>cP12</td>
<td>Prwma</td>
<td>62</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>Co$_3$Sb$_2$</td>
<td>mP12</td>
<td>P15$_3$ac1</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>CrSb$_2$</td>
<td>hP6</td>
<td>P6$_3$m22</td>
<td>152</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>Cu$_3$Sb</td>
<td>tP6</td>
<td>P6$_3$/mmm</td>
<td>128</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>CuAl$_2$</td>
<td>hH2</td>
<td>W3/m nause</td>
<td>140</td>
<td>60</td>
</tr>
<tr>
<td>12</td>
<td>CuCo$_2$</td>
<td>b6</td>
<td>W3/mnm</td>
<td>132</td>
<td>29</td>
</tr>
<tr>
<td>13</td>
<td>Fe$_3$P</td>
<td>hP6</td>
<td>P-62m</td>
<td>139</td>
<td>6</td>
</tr>
<tr>
<td>14</td>
<td>Fe$_3$Sb</td>
<td>cP12</td>
<td>Fm-3m</td>
<td>208</td>
<td>18</td>
</tr>
<tr>
<td>15</td>
<td>HfGe$_2$</td>
<td>tH2</td>
<td>W3/mjamd</td>
<td>141</td>
<td>7</td>
</tr>
<tr>
<td>16</td>
<td>Ho$_2$U</td>
<td>hP3</td>
<td>P6$_3$/mmm</td>
<td>191</td>
<td>17</td>
</tr>
<tr>
<td>17</td>
<td>Ho$_3$Sb$_2$</td>
<td>cB6</td>
<td>C222</td>
<td>31</td>
<td>6</td>
</tr>
<tr>
<td>18</td>
<td>KHf$_2$</td>
<td>cH12</td>
<td>Imma</td>
<td>74</td>
<td>61</td>
</tr>
<tr>
<td>19</td>
<td>La$_2$Sb</td>
<td>hH2</td>
<td>W3/mnm</td>
<td>139</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>LaSb$_3$</td>
<td>cS24</td>
<td>Cmca</td>
<td>64</td>
<td>6</td>
</tr>
<tr>
<td>21</td>
<td>MgCu$_2$</td>
<td>cF24</td>
<td>Fd-3m</td>
<td>227</td>
<td>108</td>
</tr>
<tr>
<td>22</td>
<td>Mg$_2$Cu</td>
<td>cF24</td>
<td>Fd-3m</td>
<td>70</td>
<td>7</td>
</tr>
<tr>
<td>23</td>
<td>MgZn$_2$</td>
<td>hP12</td>
<td>P6$_3$/mmc</td>
<td>164</td>
<td>97</td>
</tr>
<tr>
<td>24</td>
<td>MoP$_2$</td>
<td>b8</td>
<td>I4/mmm</td>
<td>71</td>
<td>10</td>
</tr>
<tr>
<td>25</td>
<td>MoS$_2$</td>
<td>b8</td>
<td>I4/mnm</td>
<td>132</td>
<td>34</td>
</tr>
<tr>
<td>26</td>
<td>NdAs$_2$</td>
<td>mP12</td>
<td>P42$_1$/ac1</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>27</td>
<td>NbS$_1$</td>
<td>hP6</td>
<td>P6$_3$/mmm</td>
<td>194</td>
<td>29</td>
</tr>
<tr>
<td>28</td>
<td>Ce$_3$Ge$_2$</td>
<td>mS12</td>
<td>C22/m1</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>29</td>
<td>PbCl$_2$</td>
<td>cP12</td>
<td>Prima</td>
<td>62</td>
<td>29</td>
</tr>
<tr>
<td>30</td>
<td>ThSi$_2$</td>
<td>tH2</td>
<td>W3/mjamd</td>
<td>141</td>
<td>7</td>
</tr>
<tr>
<td>31</td>
<td>Ti$_3$Ni</td>
<td>cF24</td>
<td>Fd-3m</td>
<td>227</td>
<td>21</td>
</tr>
<tr>
<td>32</td>
<td>TiS$_2$</td>
<td>cF24</td>
<td>Fd-3m</td>
<td>70</td>
<td>6</td>
</tr>
<tr>
<td>33</td>
<td>ZrGe$_2$</td>
<td>cS12</td>
<td>Cmcm</td>
<td>65</td>
<td>9</td>
</tr>
<tr>
<td>34</td>
<td>ZrS$_2$</td>
<td>cC12</td>
<td>Cmcm</td>
<td>63</td>
<td>16</td>
</tr>
</tbody>
</table>

| Total | 340 |
From the International Tables for Crystallography, Vol. A1 [2.3], the maximal $t$-subgroups (i.e. with the same translations) of P6/mmm (191) are the following eight space groups: P $\bar{6}$ 2m (189), P $\bar{6}$ m2 (187), P6mm (183), P622 (177), P6/m11 (175), P$\bar{3}$m1 (164), P$\bar{3}$1m (162), and Pmmm (65). Also, there are four maximal $k$-subgroups (i.e. of the same class), P6$/\overline{3}$mmc (194), P6$/\overline{3}$mcm (193), P6/mcc (192), and P6/mmm (191) itself. The hierarchy of the space groups and the corresponding 12 subgroups is shown in Figure 2.5(a) and 2.5(b). Except for a structure type, CrSi$_2$, other 5 structure types are belonged as the subgroup to the aristotypes, AlB$_2$ and Hg$_2$U.

![Diagram](image)

**Figure 2.5.** The relationship of symmetry groups of a hexagonal system; (a) the maximal $t$-subgroups and (b) maximal $k$-subgroups of a space group P6/mmm. Note that P$\bar{3}$1m (162) and P$\bar{3}$m1 (164) belong to the trigonal system, and Cmmm (65) to orthorhombic system, in (a).

As an example, consider two structure types, AlB$_2$ (P6/mmm) and CaIn$_2$ (P6$/\overline{3}$mmc), that are symmetrically related, in which both are belonged to the hexagonal system, meaning that they have a group-subgroup relation (Figure 2.5). P6/mmm denotes
the short symbol expression of P6/m2/m2/m, i.e. a primitive lattice which includes six- and two-fold rotation axes and reflection (mirror) planes perpendicular to the axes. Similarly, P63/mmc denotes the short symbol of P63/m2/m2/c, a primitive lattice which includes a screw transformation (63) along with two-fold rotation, reflection (2/m) and glide (2/c) planes. As the symmetry is reduced by a \textit{klassengleiche} transformation, the index \( n = 2 \) is calculated according to the equation (2.1) which is shown above (since \( M(G_1)=3, M(G_2)=6, \) and \( P(G_1)=P(G_2)=24 \)). Both P6/mmm and P63/mmc belong to the same crystal class of 6/mmm (i.e., \( k \)-subgroup). This relation is thereby expressed as follows:

\[
\text{AlB}_2 \text{ (P6/mmm)} \rightarrow k2 \rightarrow \text{CaIn}_2 \text{ (P63/mmc)}
\]

As mentioned, the space group information of both structures is not enough to completely understand the structural relations, and thus the atomic coordinates should be taken into account. For the structure type AlB\(_2\) (P6/mmm, 191), the basis vectors of Al (Wyckoff site 1a, 6/mmm) and B (Wyckoff site 2d, \( \bar{6} \text{m}2 \)), and for CaIn\(_2\) (P63/mmc, 194) structure, that of Ca (2b, \( \bar{6} \text{m}2 \)) and In (4f, 3m.) are known from the references 2.2 and 2.4. The unit cell of CaIn\(_2\) is extended along the \( z \)-axis (i.e. \( c'=2c \)) and has a distorted atomic positions by the alternating displacement of indium atoms (i.e., puckering of indium hexagonal layers) against that of the AlB\(_2\) prototype [2.6]. The detailed information on the atomic arrangement is given as follows.

- Atomic positions of AlB\(_2\) structure type:
  
  Al (1a) \( 0, 0, 0 \);
  
  B (2d) \( 1/3, 2/3, 1/2; \ 2/3, 1/3, 1/2 \)

- Atomic positions of CaIn\(_2\) structure type:
Ca (2b)  \( 0, 0, 1/4; 0, 0, 3/4; \)

In (4f)  \( 1/3, 2/3, z; 2/3, 1/3, z+1/2; 2/3, 1/3, z; 1/3, 2/3, z+1/2 \)

From the comparison of the relative atomic positions between the two structures, it is recognized that the origin and coordinates need to be shifted to describe them on a same setting of the coordinate. The symmetrical relations of AlB\(_2\) and CaIn\(_2\) structure types are summarized in Figure 2.6.

*Figure 2.6.* The symmetry relations between two structure prototypes, AlB\(_2\) and CaIn\(_2\), and the symmetry reduction; where, \(k2\) denotes the symmetry reduction to a subgroup of the same class with the transformation index of 2. The information as to the matrix and origin shift came from the reference 2.3. The figures on the crystal structures of AlB\(_2\) and CaIn\(_2\) came from the reference 2.6.

According to the Figure 2.4, the maximal subgroups, among the space groups of 34 AB\(_2\) structure types we consider, of AlB\(_2\) (and also Hg\(_2\)U) structure are four space
group types for six structures, i.e. P6/mmc (CaIn$_2$, MgZn$_2$, and Ni:In), P6 2m (Fe$_2$P), P 3 m1 (Cd:Ce), and Cmmm (ZrGa$_2$). The group-subgroup relations of these structures are shown in the Figure 2.7. Since Cd:Ce and ZrGa$_2$ structures belong to trigonal (a subsystem of hexagonal system) and orthorhombic system, respectively, the distance from the supergroup (P6/mmm) to these two structure types is described as longer than that of other subgroups in the hexagonal system.

![Figure 2.7. The relations between a (super) space group P6/mmm (for the structure types, AlB$_2$ and Hg:U) and their maximal subgroups.](image)

In the same way, the symmetry relations starting from the cubic system, CaF$_2$ (Fm$ar{3}$m, 225) can be investigated by using information from the International Table [2.3]. As shown in the Figure 2.8, there exist five maximal $t$-subgroups and two maximal $k$-subgroups. Among 34 structure types, only four structure types with I4/mmm symmetry (139), i.e. CaC$_3$, CuZr$_2$, LaSb and MoSi$_2$, correspond to the maximal subgroups. According to the International Table [2.3], the index is 3 (the orders, $P$(Fm$ar{3}$m) / (I4/mmm) = 48/16 = 3; see the Table 2.1).
Figure 2.8. The relationship of symmetry groups of a cubic system; (a) the maximal $t$-subgroups and (b) maximal $k$-subgroups of a space group $\text{Fm} \bar{3} \text{m}$. Note that $\text{I}_4/\text{mmm}$ (139) and $\text{R} \bar{3} \text{m}$ (166) belong to the tetragonal system and trigonal subsystem of hexagonal system, respectively, in (a).

Since the symmetry $\text{I}_4/\text{mmm}$ belongs to the maximal $t$-subgroup of $\text{Fm} \bar{3} \text{m}$, the relations can be described as shown in the Figure 2.9.

Figure 2.9. The relations between a (super) space group $\text{P}6/\text{mmm}$ (for the structure types, $\text{AlB}_2$ and $\text{Hg}_2\text{U}$) and its maximal $t$-subgroup $\text{I}_4/\text{mmm}$ (for the structure types, $\text{CaC}_2$, $\text{CuZr}_2$, $\text{La}_2\text{Sb}$, and $\text{MoSi}_2$) with its index equals 3.
Figure 2.10. The symmetry relations between structures are refined by organizing group-subgroup relationships from the two starting structure types (CaF$_2$ and AlB$_2$/Hg$_2$U) correspond to the cubic and hexagonal systems, respectively (compare with Figure 2.3). This figure is the combined results of the Figure 2.7 and 2.9 above, and is also a portion of the diagram of Figure 2.11. Note that in the hexagonal system, a space group P-3m1 (Cd$_2$Ce) is a common subgroup of two (super) space groups, i.e. P6$_3$/mmc (CaIn$_2$/MgZn$_2$/Ni$_2$In) and P6/mmm (AlB$_2$/Hg$_2$U), respectively.

This kind of investigation for the structural relations between different structure types regarding the symmetry is systematically implemented upon all 34 structure types. The supergroup-subgroup relationships are not limited within the same crystal system. The whole list of subgroup is investigated for all 34 structures. If there are no direct group-subgroup relations, then the intermediate space group(s) which connects the two space groups should also be searched. The whole result was
shown as a network of structural relations in Figure 2.11. One thing which should be noted for this diagram is that the connection between different structures cannot be directly related to the physical and chemical reasoning.

**Figure 2.11.** The symmetrical hierarchy of crystal-structure types of AB$_2$ system. The structural relations of 34 AB$_2$ prototypes have been derived in a form of family tree which present the supergroup-maximal subgroup relations.
References


CHAPTER 3
DESIGN RULES OF CRYSTAL STRUCTURES USING RECURSIVE
PARTITIONING METHODS

3.1. Introduction

3.1.1. The mapping of materials information
The mapping of materials information was briefly described in the chapter 1. Periodic Table shows that how the fundamental regularities hidden in approximately 100 chemical elements can be disclosed by systematically organizing the known data on the elements. In a similar manner, structure map – a representative classification scheme that correlates the crystal structure of compounds with the physical, or phenomenological, factors influencing their formability and structural stability - has fulfilled its role as an explanatory and predictive guideline for the novel materials design. In this chapter, we extend the current approach of structure-mapping that uses a limited number of two- or three-parameter coordinates towards multi-dimensional space, so called a “multivariate structure map.” The underlying connection between the stability of the crystal structure of binary compounds and some physical parameters derived from atomic properties of their constituents is uncovered through a data classification method, the partitioning of multidimensional parameter space. In this new type of structure mapping, the result of the sequential partitioning is represented with a tree structure that provides a set of IF-THEN rules for the estimation of the possible stable
structure of new compounds. We present that the regularities in this multivariate structure map can be used to identify stable structures of binary compounds not yet known, combined with first principles calculations. The procedure of our approach is described as the example using 7 physical parameters previously developed for the structure-map coordinates and crystal-structure data of 840 AB$_2$ binary compounds that were collected from the Linus Pauling File.

This chapter deals with the prediction of crystal structures by using the integration of the multivariate-data classification and the energy calculations based on density-functional theory. Figure 3.1 (a) and (b) show the example of two-dimensional mapping of the crystal structure information for chemical elements and binary compounds, respectively. When color-coding the Periodic Table according to the crystal structure, a clear pattern among the elements is recognized. That is, it is noted that most of the metallic elements in the same column (same group number) have the same crystal structure. The exception of some elements from this trend, e.g. Fe and Co, is thought to be due to the magnetic effect; according to the calculations by Paxton et al. [3.10], non-magnetic Fe and Co have hexagonal close-packed (HCP) and face-centered cubic (FCC) structure, respectively, and thus these hypothetical elements continue to follow the structural trend along with their same-column elements. Similarly in spirit to the Periodic Table, structure map can retain its graphical representation in the two-dimensional diagram by applying phenomenological parameters, e.g. Pettifor’s Mendeleev number. Figure 3.1 (b) shows a structure map for binary AB-type compounds constructed by P. Villars et al. [3.11] using the Mendeleev number††. For each chemical element a Mendeleev

†† In the original paper [3.11], the diagram is called as “atomic-environment-type map” instead of “structure map.”
number is assigned, and the respective elements corresponding to A or B component get the position on the x- and y-coordinate according to the ordering number. In this diagram, the compounds with the same atomic configuration are grouped together forming a structure domain (i.e., same colored region on the map). As Mendeleev’s Periodic Table and the prediction of unknown elements from it had preceded modern theory of atomic structure which later provided the understanding of his periodic law, likewise structure map has been considered as an instrumental tool for the prediction of the structure of compounds although it often tolerates the lack of underpinning principles. A theoretical explanation on the phenomenological structure map is provided by D. G. Pettifor [3.8].

In this chapter, we turn our eyes to the physical-coordinate maps. When using physical coordinates for the structure mapping, the underlying postulate is that the physical parameters used, which segregate dissimilar structure domains, would reflect the principles governing crystal structure of the compounds. Specifically, we focus on a classification scheme of crystal structure data which allows the use of multiple parameters. As mentioned, if more than four parameters are required to represent the principles of structural stability, the current type of structure maps cannot describe them entirely complete manner. New methods for the handling of multivariate data are demanded, which in principle are not dependent on the visual recognition to split different structure domains in a two- or three-dimensional layout. Figure 3.1 (c) and (d) schematically illustrate our approach. We introduce a multivariate data classification method, the partitioning of data space, in which the multidimensional space of atomic and physical parameters (APPs) is partitioned into smaller subspaces without reducing the dimension to define different structure domains. By a series of partitioning, each compartment encloses more homogeneous
dataset, that is, the minimum number of different structure types in a domain, and becomes more informative state. At each step of partition, a hyperplane orthogonal to a coordinate is applied to the multi-axis space for the demarcation of the boundaries between dissimilar structure domains and information entropy is used as the criterion to determine the position of the hyperplane. That is, the borders of different domains are drawn where the information entropy of the entire dataset is minimized. The outcome of classification, i.e. the relationships between different crystal structures and APPs, is apparently visualized in so called “classification tree,” which shows the distinguished (structure) classes along with the (APP) criteria used. Ultimately, intractable crystal chemistry rules which transcend human inference can be disclosed from the classifiers constructed by the machine-learning algorithm. By searching the if-then rules of the tree, the possible crystal structure for a new compound not yet explored can be suggested. In addition, this information-theoretic classification rationalizes the selection of a set of key APs as the axes of structure map by using a valid evaluation process [3.12].
Figure 3.1. From the Periodic Table to multivariate structure map. (a) the trend of the crystal structure of metallic elements in the Periodic Table, (b) a structure map for AB-type compounds classified according to the atomic-environment type using Mendeleev numbers [3.11] as the axes, (c) Low-dimensional, 2-D or 3-D, structure maps can be extended to (d) multi-dimensional structure maps by using multivariate-data classification methods. The coordinates of structure maps formed with the parameters influencing the crystal structure are designed to make groups with the compounds of the same crystal structure and to separate out different structure groups.

In this study, a predictive model based on the multivariate classification tree is built with the crystal structure data of 840 AB\(_2\)-type binary compounds and 7 APPs. The stoichiometry of AB\(_2\) is highly frequently occurred in binary compound materials,
and also includes some important structure types for binary and ternary compounds such as AlB₂, CaF₂, and Laves phases. The brief description of the parameters is tabulated in Table 1. First three parameters (\(VE, \Delta X^{MB}, \Delta R^{Z_{s+p}}\)) have been introduced by P. Villars for the structure-mapping of stoichiometric binary intermetallic compounds, i.e. AB, AB₂, AB₃, and A₃B₅, etc. [3.6]. The next two parameters (\(n_{av}, \Delta X^{Pauling}\)) were selected by E. Mooser and W. B. Pearson to describe the structural stability of normal valence compounds [3.2]. The last two (\(\Delta \Phi', \Delta n_{ws}^{1/3}\)) are the parameters defined by A. R. Miedema to estimate the energy of alloy formation [3.3]. Since the three models were proposed for different type of compound systems in each case, it is of great importance to investigate that how the various parameters can be used for the classification of dissimilar systems. Since none of the parameters plays the dominant role for all different material systems, it is valuable to investigate which parameters more dominantly influence on which systems than others. It should be noted that different types of electronegativity scale were used in the three models. The physical definition of the respective electronegativities can be understood from the unit of each parameter. Pauling electronegativity, \(\Delta X^{Pauling}\), has a (energy)\(^{1/2}\) unit and Martynov-Batsanov’s scale, \(\Delta X^{MB}\), has an (energy/valence electron) unit and Miedema’s parameter, \(\Delta \Phi'\), has a (volt) unit like a work function. After Villars [3.13], these seven parameters we use can be categorized into four different sub-groups, i.e. atomic-size factor (for \(\Delta R^{Z_{s+p}}\), valence-electron factor (for \(VE\)), atomic-number factor (for \(n_{av}\)), and electrochemical factor (for \(\Delta X^{MB}, \Delta X^{Pauling}, \Delta \Phi', \text{ and } \Delta n_{ws}^{1/3}\)) groups. From the classification tree at the next section, it will be shown that different parameters are variably used as a criterion throughout the partition steps under the principle of minimum information entropy.
Table 3.1. Atomic and physical parameters for the crystal structure prediction [3.2, 3.3, 3.6].

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>VE</td>
<td>Average number of valence electrons per atom</td>
<td>Villars</td>
</tr>
<tr>
<td>$\Delta X_{MB}$</td>
<td>Weighted difference of Martynov-Batsanov electronegativities</td>
<td></td>
</tr>
<tr>
<td>$\Delta R_{Zs+p}$</td>
<td>Weighted difference of Zunger pseudopotential radii sum</td>
<td>Mooser &amp; Pearson</td>
</tr>
<tr>
<td>$n_{av}$</td>
<td>Average principal quantum number</td>
<td>Miedema</td>
</tr>
<tr>
<td>$\Delta X_{Pauling}$</td>
<td>Pauling electronegativity difference</td>
<td></td>
</tr>
<tr>
<td>$\Delta \Phi^*$</td>
<td>Chemical potential difference for electronic charges</td>
<td></td>
</tr>
<tr>
<td>$\Delta n_{WS}^{1/3}$</td>
<td>Electron density difference in Wigner-Seitz atomic cell</td>
<td></td>
</tr>
</tbody>
</table>

The two-dimensional scatter plot of Figure 3.2 shows the distribution of $\text{AB}_2$ compounds with different crystal structures in the coordinate systems of seven physical parameters. It appears that most of the domains seem to include overlapping regions in the multivariate space and do not clearly show the separation among different structure domains. Each plot reflects, however, the two-dimensional projection of multi-dimensional parameter space, and dissimilar structures occupy distinguishable areas which can be considered as the regularities of structural stability. How to recognize and define the respective structure domains by using information-entropy based classification method called “recursive partitioning” is explained below and also at the “Methods” section.
Figure 3.2. Scatter plot of seven physical parameters. The coordinates of each plot are atomic or electronic parameters for the compounds originated from their constituent elements. The color-coded points on the plot represent 840 AB$_2$-type compounds that are classified into 34 classes according to their crystal-structure type. Each sub-plot indicates the two dimensional view of the seven dimensional parameter space, which corresponds to a kind of structure map. For the definition of each parameter, see the description in Table 3.1.

3.1.2. Recursive partitioning and classification tree

The classification of data with the partitioning of variable space and the resultant classification tree is schematically illustrated in Figure 3.3. Shannon’s information entropy, H, is used as a measure of the uncertainty of data which needs to be reduced during the classification [3.14]; the details will be covered at the “Methods”
section. Suppose a simple system in which only two variables are used. Before the partitioning of this two-dimensional parameter space, ten different components categorized into three classes, i.e. A, B, and C, are placed in a single bin (State 1) and the entropy of the system is in a maximum state. As the components are subdivided into smaller bins by adopting two hyperplanes, the entropy of entire system becomes minimized (State 2). Since the possible position of hyperplanes is restricted in a way to minimize the entropy, this binning process [3.15] can be considered as a series of procedure to determine the boundaries among different classes which maximize the reduction of the entropy of the entire bins. The partitioning steps of multi-dimensional space are in general represented with a tree structure as the result (Figure 3.3b). This tree-structured diagram provides the information on the classified groups and the splitting criteria for the boundaries between them. In the multi-dimensional space partitioning, a set of splitting condition can be considered as the addresses to reach at each compartment.

![Diagram](image)

**Figure 3.3.** (a) Information entropy change ($\Delta H$) by the partitioning of the parameter space, and (b) a tree structure for the graphical representation of the classification result; A, B, and C denote different classes, namely, the compounds with dissimilar structure types. R is a root node, i.e. the state 1 before the partitioning.
If once the classification of the given data is completed, the membership of a new component can be assigned by simply investigating where the component is located in the subdivided parameter space. As pointed out previously, when appropriate parameters are selected for a structure map, the compounds of the same structure are grouped together and thus each structure type occupies its distinct region in the map. For actual crystal structure data, however, these dissimilar structure domains (e.g. A, B, C, and D in Figure 3.4 left) often include overlapping regions. As D. G. Pettifor indicated [3.16], if so, how can we decide the class membership in which an unknown member is involved, under the existence of “blurry” boundaries of different domains? The combination of the recursive partitioning with the “look-up the neighbors” view as referred by D. Morgan [3.17] offers the feasibility of applying this multivariate classification for the prediction of crystal structure. At first, in the Figure 3.4, the parameter space is split into smaller bins in which the entropy is minimized. Some bins are completely homogeneous, i.e. only one structure-type domain is assigned (e.g. region 3 in Figure 3.4 right side) and other overlapping domains are existed as well (e.g. region 1 and 2 in Figure 3.4 right side). By matching the position of a new compound of which the structure is not yet determined on the map, the possible crystal structure(s) can be estimated. For instance, if a compound is positioned in the region 3, structure D would be considered as the most likely stable form of the compound. In cases of the region 1 and 2, more than one structure is nominated as the candidate. That is, A, B, or C for the region 1 and A or C for the region 2 would be suggested as the possible structures of the compound, respectively.
Figure 3.4. Illustration of the partitioning of parameter space: (a) in a parameter space, the compounds with the same crystal structure are grouped together forming a cluster referred to as a “structure domain” (A, B, C, and D are different structure domain regions); (b) by partitioning the parameter space, the region encompassed by each structure domain is defined by a range of two parameters 1 and 2. See the text for the explanation on the prediction of the crystal structure of three hypothetical compounds (at the points 1, 2, and 3) by the partitioning of parameter space. Note that structure domains, A, B, and C have overlapping regions; in actual crystal structure maps, there often exist many structure domains that are overlapped each other.

The idea of partitioning space in Figure 3.4 can be extended, in the same way in principle, to the higher dimensional parameter space. Let us imagine an n-dimensional box (e.g. n=7 in this study) which contains a few hundred color balls (suppose a set of compounds color-coded by their structure types, thus same colored balls imply the same structure type); a compound described by seven physical parameters can be considered as a ball which is positioned somewhere in the seven-dimensional parameter space. Recursive partitioning is then a series of processes to separate the balls in the box with several flat plates (i.e. linear classification criteria) in order to sort out more same colored balls into a same compartment (forming a more informative dataset).
3.2. Methods

3.2.1. Construction of data set
To build a classification model, the crystal structure data of intermetallic compounds were collected from LPF. Out of 103 elements (atomic number 1-103) in the Periodic Table, 68 elements were taken into account after excluding hydrogen, chalcogen, halogen, noble-gas group elements, and Pm, Eu, Tb, Yb, Pa, Np, Am, Cm, Bk, Cf, Es, Fm, Md, No, and Lr in the lanthanides/actinides. In the LPF version 1.0, AB$_2$ binary compounds formed by the selected 68 elements have 973 entries which can be categorized into 109 structure-type classes. Among them, 840 compounds (86.3%) are represented by main 34 structure types which have more than six entries, and the rest of 133 compounds shares 75 structure types. That is, the 75 structure types, so called minor structure types, in which each structure type has less than five compounds as the members, were excluded from the dataset we considered. In addition, the data for the compounds prepared at high temperature and pressure, low temperature, stabilized by impurity, metastable and/or polymorphic states were not used for this study.

3.2.2. Selection of atomic parameters
The correct selection of atomic, physical parameter (APP) sets is critical to construct a good predictive model. In this study, we chose seven parameters suggested by Mooser & Pearson, Miedema, and Villars [3.2, 3.3, 3.6, 3.13] in their structure-map models. By selecting the parameters well-known as the factors responsible for structural stability, we intended to concentrate on showing how the predictive classification of crystal structure data is established by using data mining approach.
Since the parameters selected are variably used and their relative contribution is simultaneously evaluated during the classification, multiple parameters can be tested without any dimensional limitation of structure mapping. The ranges of some APP values for the AB$_2$–type structures are compared in Figure 3.6. This kind of plots can be considered as one-dimensional structure maps. As shown, there exists broad range of overlapping regions of APP values among dissimilar structure types. It implies that one-parameter plots using these four parameters are not so effective for the assessment of the parameters used, as well as for the differentiation of dissimilar structure types. By using the combination of appropriate number of multiple parameters, instead, the resolution of the discrimination among the structures can be improved, and thus the selection of effective APPs has become the critical part of structure mapping approach.
Figure 3.6. One-dimensional structure mapping. For different structure types of AB₂ compounds, the range of the values of some parameters was represented using LPF data. Square symbols indicate the existing data of the compounds with structure information in the LPF.

3.2.3. Information entropy

Information entropy was for the first time introduced by C. E. Shannon for the application to telecommunication systems [3.14]. Since the new concept was established, the wide applicability of the theory has been found in diverse fields. In this study, we describe the information entropy function as a splitting criterion for the classification of crystal structure data. Information entropy, H, of a dataset is a measure of the uncertainty and mathematically defined as [3.14, 3.26]:

\[
H = - \sum_{i=1}^{n} p(x_i) \log p(x_i)
\]  

(3.1)

Where, \( p(x_i) \geq 0 \), \( \sum_{i=1}^{n} p(x_i) = 1 \) \( (i = 1, 2, \ldots, n) \)

In the above equation (1), \( p(x_i) \) is the frequency of the occurrence of a particular structure type, \( x_i \), in the given dataset.
Out of 840 compounds, for instance, there are 45 AlB$_2$, 16 CaC$_2$, 21 CaF$_2$, ..., 15 ZrSi$_2$ type structures in the LPF database. At the root node (i.e., before partitioning), the entropy of the data, $H_{\text{Root}}$, is:

$$H_{\text{Root}} = - \sum p(x_i) \log_2 p(x_i)$$

$$= -(p(\text{AlB}_2) \log_2 p(\text{AlB}_2) + p(\text{CaC}_2) \log_2 p(\text{CaC}_2) + p(\text{CaF}_2) \log_2 p(\text{CaF}_2) + \ldots$$

$$+ p(\text{ZrSi}_2) \log_2 p(\text{ZrSi}_2))$$

$$= -(\frac{45}{840} \log_2 (\frac{45}{840}) + \frac{16}{840} \log_2 (\frac{16}{840}) + \frac{21}{840} \log_2 (\frac{21}{840}) + \ldots$$

$$+ \frac{15}{840} \log_2 (\frac{15}{840}))$$

$$= 4.3751 \text{ bits}$$

### 3.2.4. Classification tree: Partitioning of multidimensional space

The aim of the classification is to reduce the uncertainty of the dataset, and thus to extract useful information from the data. In practice, the classification in this study, more specifically *partitioning*, is to consecutively find an APP value which minimizes the information entropy of the entire dataset by partitioning the data with the cutoff value. This corresponds to a series of consecutive processes to search the position of a hyperplane which bisects a portion of the APP space in order to minimize the entropy. The classification tree which is built as the result of the partitioning connects the respective compounds to the crystal structures with the APPs, and for extracting useful knowledge from the causal linkage between the two, i.e. APPs and crystal structures. Initially, the respective compounds, $S\{s_1, s_2, \ldots, s_l\}$, in a dataset (here, $l=840$), in which each compound is characterized by the APPs, $V=\{v_1, v_2, \ldots, v_m\}$ (here, $m=7$), are connected to one of the classes (i.e., structure types), $C\{c_1, c_2, \ldots, c_n\}$ (here, $n=34$). Then, the entropy criterion applied to the partitioning is for:
\[ x_i(m) = \arg\max_{x_i} (p_m(x_i)) \]

Where, \( x(m) \) is a major class (i.e., structure type) in a node \( m \) and \( p_m(x_i) \) is a probability of the class \( x_i \) in the node \( m \), and “argmax” denotes the “argument of the maximum”. Namely, as the classification tree grows, the homogeneity of each node is maximized (on the contrary, the entropy of the tree is minimized at the terminal nodes). For example, if there is only one class, then the maximum value \( p_m(x_i) = 1 \).

When this classification as a mapping (M) process, \( M: S(V) \rightarrow C \), is correctly performed, the uncovered regularities in the data are developed as sets of if-then rules.

Starting from the root, or single-group node, the compounds are divided into two offspring subgroups according to the “cutting” value of the APPs. These numeric constraint values which make best bisection are achieved by maximizing the entropy (H), before and after a partition step. Thus, the information gain (IG), \( \Delta H \), i.e. the goodness-of-split which can be achieved by a partition step, is defined as:

\[
\Delta H = H_{\text{ascendant}} - (p_1 H_{\text{descendant,1}} + p_2 H_{\text{descendant,2}})
\]

(3.2)

Where, \( H \) is the information entropy of the data at a given level of the tree defined by equation (3.1); \( p_1 \) and \( p_2 \) are the fractions of each descendant node 1 and 2, respectively, and thus here \( p_1 + p_2 = 1 \).
Although the grouping turns into more homogeneous subsets as the partition proceeds, the splitting of the tree should be stopped at a certain level to prevent the overfitting. That is, when no improvement of the model performance is observed, the bifurcation step is terminated [3.27]. This pruning procedure is highly significant not only the predictability of a tree model is evaluated during the pruning but also it is directly related with the number of the structure types suggested by the classification model and also with the computation time for the ab initio calculations. The pruning optimization is implemented until the prediction reliability is not improved during the validation test. We assessed our model by a random-sampling-based cross validation method; i.e. a part of the compound data (i.e. one-tenth of the data) is randomly selected and set aside for the test from the entire 840 compounds list. Then, a classification tree is formed with the remainder called a training dataset. Finally, we check out, using the test data, whether the predictive model, i.e. classification tree, correctly suggests the possible structure type. This procedure is repeated with ten times of random sampling. The validation process is carried out at various tree levels, and the partitioning is stopped at the level of which the prediction error rate becomes the lowest. Here, the classification error is calculated as follows:

$$\text{Classification error} = \frac{\text{the number of misclassified samples}}{\text{total number of samples}} \times 100\% \quad (3.3)$$

Thus, if there are 84 test datasets are misclassified among the 840 test data by the 10-fold cross validation, the classification error is \((84/840) \times 100 = 10\%\) and the predictability of the classification model is 90 %.

Once the prediction model of if-then rules is constructed, the crystal structure of a new compound can be estimated by simply applying to the model. Although the
grouping turns into more homogeneous subsets as the partition proceeds, the splitting of the tree should be stopped at a certain level to prevent the overfitting. That is, when no improvement of the model performance is observed, the bifurcation step is terminated [3.27]. This pruning procedure is highly significant not only the predictability of a tree model is evaluated during the pruning but also it is directly related with the number of the structure types suggested by the classification model and also with the computation time for the ab initio calculations. The pruning optimization is implemented until the prediction reliability is not improved during the validation test. We assessed our model by a random-sampling-based cross validation method; i.e. a part of the compound data (i.e. one-tenth of the data) is randomly selected and set aside for the test from the entire 840 compounds list. Then, a classification tree is formed with the remainder called a training dataset. Finally, we check out, using the test data, whether the predictive model, i.e. classification tree, correctly suggests the possible structure type. This procedure is repeated with ten times of random sampling. The validation process is carried out at various tree levels, and the partitioning is stopped at the level of which the prediction error rate becomes the lowest. Once the prediction model of if-then rules is constructed, the crystal structure of a new compound can be estimated by simply applying to the model.

3.3. Results and discussion

The procedure to construct a classification tree comprises three main steps: 1) the selection of a set of appropriate APPs which correspond to the coordinates of the chemical space of compounds, 2) the preparation of an input dataset, i.e. with crystal
structure information and APP values, and 3) the classification of the dataset by recursive partitioning, followed by the model evaluation – including the assessment on the selection of APPs. In this study, 840 AB₂-type compounds categorized into 34 structure types in the Linus Pauling File (LPF) database [3.18] were classified by using seven APPs originated from three different structure-map models. The result of the classification was shown in Figure 3.7. Each node denotes the splitting condition for achieving the minimum entropy at the given depth of the tree by separating the compounds of different structures. Before the classification, the 840 compounds of 34 structure types are contained in the root node (®). At each step of the partition, a specific parameter (or attribute) subdivides the compounds into two sub-groups reducing the information entropy of the system. After 28 partition steps, the number of possible structure types for each branch is reduced to one to ten entries; in fact, first the tree structure is overgrown and then pruned to obtain a best sub-tree that provides the highest prediction efficiency. The performance of the tree is evaluated by a typical cross-validation procedure. That is, the data are divided into two sub-datasets, i.e. training and test data. Then a model built with the training data is assessed using test data. This validation process continues until the test of whole dataset is completed. According to the cross validation of the model we constructed, 92 % of the candidate lists suggested correct structure types. This is very impressive result in that our model was primarily aiming at the illustration of the multivariate classification approach rather than the establishment of the best prediction model. To improve the efficiency of the prediction model, one can test various APPs and then reconstruct the model with an optimized set of the parameters highly contributing to the stable structure formation. It should be noticed that all the APPs used are not equally important for the classification of different crystal structure types [3.12]. Most of the compounds in this example could
be classified with only two to four parameters, not all the seven parameters; principal quantum number ($n_{av}$) was not used for the classification tree in the Figure 3.7. Some detailed description on the procedure for building the classification tree was shown at the Methods section.

In Figure 3.8, some portions of the classification tree, i.e. (a)-(h) of Figure 3.7, were described on the atomic size ($\Delta R_{z+p}$)-valence electron number (VE) plane. It clearly shows that how the growth of the classification tree is related with the partitioning steps of parameter space. The regions occupied by compounds (data points) are subdivided into sub-regions by selecting a particular APP value which makes the entropy minimum at each partitioning step. Note that the splitting hyperplane is orthogonal to the attribute axis selected. It can be seen that the domains of the compounds with AlB$_2$ structure (red-colored dots) are redefined as the partitioning proceeds. As mentioned, structure map has been a practically effective approach which unveils the relationships between the crystal structure of compounds and the relevant properties from their constituent elements. The regularities achieved by mapping inorganic compounds to their structure classes with the APP expressions become the rules of crystal chemistry. Given a stoichiometry, one can estimate the structural stability of a particular compound and achieve a guideline for the development of new materials. The tree-structured diagram, as a multivariate structure map, shows the pathways to achieve specific crystal structures which are restricted by the interaction of the relevant APPs. If once the classification tree of if {sets of APP values}-then {structure types} rules is generated, the prediction of the crystal structure of unknown compounds can be carried out by tracking the route in the tree structure according to the criteria given at each node. Consequently, one structure type among the nominated would be the most stable structure for the
specific compound. This “screening” process to find the possible structure of a compound extensively reduces the search region of ab initio calculations. To estimate the stable structure of a compound for a given composition, one can either investigate all the possible atomic configuration of the constituent elements finding a structure with the lowest energy state or employ a collection of empirical principles related to the structural stability of compounds. Currently, it seems that there is a complementary compromise between these two approaches [3.19-3.25]. By integrating two stand-alone methods, computational strategy for a wide range of materials problems can become a practically more feasible task. For example, when tested with all 840 compounds in our study, one to ten candidates are selected for a test compound as the possible stable structure type; in average, five structure types are suggested during the cross-validation test. Considering the fact that we have used 34 structure types, the load of the quantum mechanical computation is reduced to approximately 3-30% of the original load.
Figure 3.7. A classification tree for the crystal structures of AB$_2$ compounds. By 28 times of recursive partitioning, this tree structure with 29 if-then classification rules was achieved. At the terminal nodes, a few – one to ten – crystal structures (prototypes) are suggested as the possible stable structures. Some parts of the tree were highlighted with dot-line boxes to explain the partitioning of parameter space.
The pathways indicated by red arrows are juxtaposed with the corresponding parameter-space regions, i.e. grey-colored area in the plots (a)-(h) in Figure 3.8. Note that only two to four APPs are variably used in the series of *if-then* rules for each branch; in these most cases, the combination of atomic size, electrochemical, and valence-electron factors (after Villars, Ref. 3.13) are used for the partitioning.
**Figure 3.8.** Partitioning of chemical space that includes AlB$_2$ structure domain. Each plot shows a series of partition steps (a)-(h) shown in the classification tree through the 2-D projection of $\Delta R_{\text{sp}}$-VE plane. Data points correspond to the 840 AB$_2$ compounds used to build the classification tree of Figure 3.7; among them, red-colored dots indicate the compounds with the AlB$_2$-type structure. Gray regions highlight the newly identified part by each partitioning step at the classification tree (compare with the nodes indicated by red arrows in Figure 3.7).
<table>
<thead>
<tr>
<th>Group</th>
<th>If-then rules for the prediction of possible crystal structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is MgF2.</td>
</tr>
<tr>
<td>2</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is MgCu2.</td>
</tr>
<tr>
<td>3</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is CuFe2.</td>
</tr>
<tr>
<td>4</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is FeMg.</td>
</tr>
<tr>
<td>5</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is MgFe.</td>
</tr>
<tr>
<td>6</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is FeCu.</td>
</tr>
<tr>
<td>7</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is MgFe2.</td>
</tr>
<tr>
<td>8</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is FeMg2.</td>
</tr>
<tr>
<td>9</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is MgCu.</td>
</tr>
<tr>
<td>10</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is FeCu2.</td>
</tr>
<tr>
<td>11</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is MgFeCu.</td>
</tr>
<tr>
<td>12</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is MgCuFe.</td>
</tr>
<tr>
<td>13</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is FeMgCu.</td>
</tr>
<tr>
<td>14</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is FeCuMg.</td>
</tr>
<tr>
<td>15</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is MgFeCu2.</td>
</tr>
<tr>
<td>16</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is FeMgCu2.</td>
</tr>
<tr>
<td>17</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is MgFeCuMg.</td>
</tr>
<tr>
<td>18</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is FeMgCuFe.</td>
</tr>
<tr>
<td>19</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is MgFeCuFe.</td>
</tr>
<tr>
<td>20</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is MgFeCuMgFe.</td>
</tr>
<tr>
<td>21</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is FeMgCuFeMg.</td>
</tr>
<tr>
<td>22</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is MgFeCuMgFeMg.</td>
</tr>
<tr>
<td>23</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is MgFeCuFeMgFe.</td>
</tr>
<tr>
<td>24</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is MgFeCuFeMgFeMg.</td>
</tr>
<tr>
<td>25</td>
<td>If A, B, C are Mg, Fe, Cu, then the possible structure is MgFeCuFeMgFeMgFe.</td>
</tr>
</tbody>
</table>
If-then rules which correspond to the classification tree are tabulated in the Table 3.2. In this chapter, we have demonstrated that how the unknown crystal structure of a new binary compound can be estimated from a multivariate classification of the existing crystal structure data by correlating them with first principles calculations. Using the structure data of 840 AB$_2$ binary compounds and some atomic and physical parameters (APPs) which influence the structural stability, we showed that our new type of multi-dimensional structure mapping can be applied as a rational "design" tool to sift and sort the possible structures of a new compound otherwise laborious computations might be required. In particular, this machine-learning algorithm plays the important role as an effective research scheme which complements the pure human-based learning in that not only does it accelerate the learning process but further it may disclose some hidden rules not yet detected by human intuition and experience [3.12].

As P. Villars has already pointed out [3.6], the resolution of the prediction which is based on the existing data is inherently restricted within the structure types that were taken into account. In other words, since the range of predictable structures is limited within the structure types one considered to construct the model, other minor structure types ignored due to their low occurrence or new structures not yet found cannot be readily predicted in the scope of the given model. However, considering the accelerating speed of the data accumulation from experimental and theoretical implementations in these days, more challenging problem, namely bottleneck, of future materials research will be encountered from the stagnation in extracting useful knowledge from the huge volume of data. As a whole, the application of data mining and informatics is inevitable. Specifically, the crystal-chemical rules derived from a machine-learning scheme, i.e. if-then rules formulated
with atomic properties and the known information on the crystal structure of compounds will provide the comprehensive understanding on the structural stability of compound materials. In the present study, we applied this method for the binary compounds. However, this approach can be equally extended to the multinary compound systems by exploring more effective factors which govern crystal structures.
References


4.1. Introduction

It has long been one of the main subjects of crystal-chemistry research that relating the underlying regularities in the crystal structure and stability of inorganic compounds to a number of factors originated mainly from atomic, physical parameters (APPs) and interactions (i.e. chemical bonding) of their constituent elements. In some groups of compounds, for instance, the ratio of atomic size (e.g., ionic-radii ratio) between their components highly influences on the coordination of the bonding and hence the crystal structure of the compounds. It is also often observed that different compounds with the same valence-electron concentration (VEC) occupy similar atomic configuration. Once if the observations on this kind of tendency can be quantitatively formulated as the design rules for novel materials, one would ultimately be able to manipulate the crystal structure, and properties, of substances by selecting a certain combination of chemical elements from the Periodic Table according to the guideline of the rules. Towards this goal, quite a number of building-up principles, i.e. “factors governing the crystal structure,” of compounds have been developed from either theoretical or semi-empirical approaches that basically connect some of the critical factors with the atomic configuration of the inorganic solids by recognizing the regularities embedded in the data [4.1-4.5]. A brief and partial list of the references on the principles of the factors influencing
The crystal structures of compound materials is summarized in Table 4.1. It is recognized that the factors shown in the Table 4.1 are the parameterizations intrinsically originated from the electronic configuration of the constituting chemical elements. The effects of the factors on the crystal structure of inorganic compounds are investigated by searching any pattern which connects the factors with several known crystal structures of compounds.

**Table 4.1.** A brief and partial list of the references on the factors governing crystal structures [4.1-4.5].

<table>
<thead>
<tr>
<th>Reference</th>
<th>Factors and Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hume-Rothery† 1926, 1934 [4.1]</td>
<td>Valency effect (average number of valence electrons per atom, e/a), size factor, the effect of the periods of the solvent and solute atoms (electrochemical factor)</td>
</tr>
<tr>
<td>Laves 1956, 1967 [4.2]</td>
<td>Geometrical factors (highest degree of space filling, symmetry, and connections), size factor, electrochemical factor, and bond factor</td>
</tr>
<tr>
<td>Engel-Brewer 1964, 1967 [4.3]</td>
<td>Electronic configurations (s- and p-electron count per atom, i.e. the number of (s+p) electrons)</td>
</tr>
</tbody>
</table>

One of the effective methods to investigate these underlying patterns is to use diagrammatic representations referred to as “structure map” [4.6-4.13]. Using two-

† In fact, Hume-Rothery introduced the rules as the empirical observation on the solid-solubility limits of metallic systems [4.1]. However, the three aspects he considered has become the foundation for the successive researches on the structural stability of inorganic compounds.
or three APPs as the coordinates of the diagram, the crystal-structural similarities of the compounds investigated are presented as the proximity between them (the distance between compounds). As the result, the compounds with the same structure are located closely each other, forming a “structure domain,” in the APP-based feature space. A key challenge in such an approach is to select the appropriate set of APPs that would reflect the regularities in the formation of stable compounds. Confronted with the practical difficulties in the selection of appropriate parameters, P. Villars has carried out a systematic search from a collection of 182 APPs to find the factors which dominantly influence the crystal structure of compounds [4.10]. Based on the results, he categorized the factors that govern crystal structures into five groups [4.5]. That is,

- Size factor (e.g., ionic, metallic, covalent, and pseudopotential radii)
- Valence-electron factor (e.g., number of valence electrons per atom)
- Electrochemical factor (e.g., electronegativity, Miedema’s chemical potential)
- Atomic-number factor (e.g., atomic number, principal quantum number)
- Angular-valence-orbital factor (after Pettifor)

It has already been known that there exist specific dominant factors preferred by different compounds. For example, intermetallic Laves phases, i.e. MgZn$_2$ (C14), MgCu$_2$ (C15), and MgNi$_2$ (C36) structure-type compounds, are well known as size-factor-dominated compounds. For the typical formation of Laves phases, the atomic sizes of the constituents are geometrically restricted in the ideal ratio of 1.225:1; in reality, however, the actual values are deviated from this ideal value, broadly varying between 1.1 and 1.6 [4.20]. Zintl compounds, e.g. Mg$_2$Pb, Mg$_2$Sn, and Mg$_2$Ge, are known as the electrochemical factor is dominating. Since, in effect, multiple
regulating factors contribute to the preferred occurrence of a certain crystal structure for different compound groups [4.2], the quantitative assessment of the relative contribution of the factors is instrumental to achieve deeper understanding on the rules hidden in the structural stability of compounds.

Data mining provides the useful tools to identify the linkage between APP-based factors and crystal structures of compounds in the existing data sets. As a major task of data mining, classification uncovers the hidden patterns and regularities from the data by grouping similar instances within the parameter space. In the initial dataset, the respective compounds, \( S\{s_1, s_2, ..., s_l\} \), in which each compound is characterized by \( m \) APP parameters, \( V = V(v_1, v_2, ..., v_m) \), \( V \) being the parameter space, are assigned to one of the classes, \( c_i \in C = \{c_1, c_2, ..., c_n\} \) (i.e., structure types), where \( \forall i \in \{1, ..., n\} \). The spatial distribution of compounds in \( m \)-dimensional APP space is described by means of the probability function, \( p(c_i|t) \), the frequency of the occurrence of a class, \( c_i \) at the classification node \( t \). A classification model is built based on the information of the probability function achieved from the dataset used. When the resulting classification as a mapping process (M), \( M: S \rightarrow C \), is correctly performed, the uncovered patterns in the data can be developed as sets of useful rules. This classification \( \Phi = \{\Phi_1, \Phi_2, ..., \Phi_x\} \) is constructed by a partition technique called “recursive partitioning,” in which the objective is to make more homogeneous groups, i.e. the minimum number of dissimilar classes in a node, with the optimum size of the partitioning. Structure-mapping method, in this respect, can be considered as a graphical classification scheme in which different chemical compounds are spatially grouped (or separated) according to their crystal-structural similarity. Since the classification of data in a structure map is represented with a limited number, mostly two or three, of coordinates, the number of APPs which can
be applied on a single structure map is restricted. Data-mining techniques allow the handling of multivariate data, and thus have a benefit in which multiple parameters can be simultaneously used for the classification. After all, the relations of APPs with the structural stability of compounds are derived as crystal-chemistry rules by quantitatively measuring the relative contribution of APPs on the classification of a given crystal structure data set.

In this chapter, we describe how a multivariate classification method based on information theory can be used for the crystal chemistry research of inorganic crystalline compounds. Partition-based classification is carried out to sort out different crystal structures, in which information entropy serves as a splitting criterion. As pointed out, information-theoretic classification outperforms two-, three-dimensional structure mappings in that many intractable rules which transcend human inference – mainly due to the inherent complexities of the system in the high dimensional space – can be disclosed by the multivariate classifiers. The outcome of the classification is visualized in a “classification tree” structure which shows a series of the splitting variables used for the partitioning (i.e. IF conditions) along with the resultant classes (i.e. THEN results). Searching the IF-THEN rules provides a short list of possible crystal structures for a given compound system, such that the crystal structure of any new compound can be estimated from the classification rules of known structural data [4.19]. The reduction of information entropy contributed by each APP during the partitioning is the quantitative measure of the correlation between APPs and crystal structures, which provides the valid evaluation criterion for the selection of key APP sets. At the next section, the concept of information entropy and the procedure of data classification by partitioning are explained in detail. We then show that how some useful crystal-chemistry rules can
be extracted from the classification process by using the structure data of 840 AB₂-
type binary compounds collected from Linus Pauling File (LPF) and seven APPs
originated from three structure-map models, i.e. Mooser-Pearson, Miedema, and
Villars model [4.6, 4.7, 4.10].
4.2. Computational details

4.2.1. Classification of data

Out of 103 elements in the Periodic Table, 68 elements were taken into account after excluding hydrogen, chalcogen, halogen, noble-gas group elements, and Pm, Eu, Tb, Yb, Pa, Np, Am, Cm, Bk, Cf, Es, Fm, Md, No, and Lr in the lanthanides/actinides. The data of the compounds prepared at high temperature and pressure, low temperature, stabilized by impurity, metastable and/or polymorphic states were not used for building up the classification model. After this pre-processing of the dataset, \( \text{AB}_2 \) binary compounds from the 68 elements include 973 compound entries in the LPF version 1.0, which can be categorized into 109 structure-types. Among them, 840 compounds (86.3\%) are represented by major 34 structure types, those have more than six entries, and the rest of 133 compounds (13.7\%) shares 75 minor structure types in which each type has less than five compounds as the members. These minor structure types were also excluded from our consideration.

Classification of data can be taken into account the partitioning of parameter space into smaller regions, minimizing the number of different classes in a region bounded by a number of linear hyperplanes. Figure 3.1 schematically shows the partitioning procedure of the data space and the corresponding tree structure. In the two-dimensional (\( X_1 \) and \( X_2 \)) coordinates system, the parameter space is split into smaller regions by straight lines (e.g. \( X_2 = p, X_1 = q, \) and \( X_1 = r \) in the Figure 4.1). In case of two-dimensional systems, the classification results can be readily presented by simply showing the partitions like typical structure maps. However, in higher dimensional space, the visualization of the partitioning is normally impossible. In this respect,
tree-type expression in which the splitting criteria and the classes are included is very effective. Here, the problem is that how to determine the position of the splitting hyperplanes, and it is specified by applying the concept of information-entropy minimization, which will be explained at the next section.

![Figure 4.1](image)

**Figure 4.1.** A schematic of (a) data classification (i.e. partitioning) in the 2-D parameter space \((X_1 \text{ and } X_2)\) and (b) its classification-tree expression ("R" indicates the root node; i.e., raw data before the classification); different symbols represent different classes of samples (e.g. different crystal structures). Where, p, q, and r represent the splitting criteria of partitioning (i.e., the position of the boundary line between different structure regions).

### 4.2.2. Classification by information entropy

In this section, we describe information entropy as the partitioning criterion for the classification of crystal structure data. The concept of information entropy was firstly introduced by C. E. Shannon for the application to telecommunication systems [4.14]. However, since the new concept has been established, the wide applicability of the theory has been found in diverse fields. For the classification of data, information entropy function, which is defined as a measure of the uncertainty
of data, is used as the criterion for partitioning data into more informative subgroups. By classifying the data in a way to minimize the average entropy of the dataset, the data become more homogeneous state; i.e. different structure types are sorted out. As the result, one can get a set of rules such as “if the condition of an attribute is smaller than or larger than a specific value, the sample would be included to this class.” Mathematically, the information entropy function, $H$, of data is defined as [4.14]:

$$H = -K \sum_{i=1}^{n} p(x_i) \log p(x_i)$$

(4.1)

Where, $p(x_i) \geq 0$, $\sum_{i=1}^{n} p(x_i) = 1$ ($i = 1, 2, ..., n$)

In the above equation (4.1), $p(x)$ is the frequency of occurrence of a particular structure type, $x_i$, in the Linus Pauling File (LPF) database, and $K$ is the constant which corresponds to a choice of the unit of measure ($K=1$ here). Entropy is the expected value of $-\log p(x_i)$. Where, the entropy is measured in “bit” unit as the logarithm of base 2 is used. Since this entropy function is a measure of the uncertainty of data, the aim of the classification is to reduce the uncertainty of the dataset so as to make it more informative form. In practice, the classification, or more specifically, partitioning is to find consecutively an APP value which minimizes the information entropy of the entire system when partitioned with the value. This corresponds to a series of processes of finding the position of a hyperplane which bisects a portion of the APP space in order to decrease the entropy.

Starting from the root node, or single group node, the compounds are subdivided into two smaller groups according to the “splitting” value of the APPs. These numeric constraints which make a best bisection at each step are to maximize the reduction of information entropy ($H$) before and after a partition step. Thus, the
goodness, $\Delta H$, i.e. the reduction of the entropy, which can be achieved by a partition step, is defined as:

$$
\Delta H = H(Y) - H(Y \mid X) = H_{\text{ascendant}} - (p_1 H_{\text{descendant,1}} + p_2 H_{\text{descendant,2}})
$$

(4.2)

Where, $H$ is the information entropy of the dataset at a given level of the tree, which is defined by the equation (4.1); $X$ and $Y$ are independent and dependent cases, respectively; $H(Y \mid X)$ is the conditional entropy of a random variable $Y$, given another random variable $X$; $p_1$ and $p_2$ are the fractions of each descendant node 1 and 2, respectively, and thus $p_1 + p_2 = 1$. This measure of goodness is called as “information gain (IG)” that is defined as:

$$
IG(Y \mid X) = \Delta H = - \sum_{i=1}^{r} p(x_i) \log p(x_i) - \sum_{i=1}^{r} p(x,y) \log p(y \mid x)
$$

(4.3)

Where, $p(x)$ is the unconditional probability of $x$, $p(y \mid x)$ is the conditional probability of $y$ given $x$, and $p(x,y)$ is the joint probability of $x$ and $y$. Here, a variable $v_i$ from $V = V(v_1, v_2, \ldots, v_m)$, where $\forall i \in \{1, \ldots, m\}$, has $l$ different values of $\{a_1, a_2, \ldots, a_l\}$ in which the value $l$ is equal to the number of instances in the dataset used. The relative contribution of the variables to each class, i.e. different crystal structures, can be quantitatively evaluated by calculating the information gain, $IG_i$, with respect to each variable which is used for the classification of particular crystal structures.

Out of 840 compounds, for an example, there are 45 AlB$_2$, 16 CaC$_2$, 21 CaF$_2$, ..., 15 ZrSi$_2$ type structures in the LPF. At the root node (i.e., before partitioning), the entropy of the data, $H_{\text{Root}}$, is:

$$
H_{\text{Root}} = - \sum p(x) \log p(x)
$$

= - \{p(\text{AlB}_2) \log p(\text{AlB}_2) + p(\text{CaC}_2) \log p(\text{CaC}_2) + p(\text{CaF}_2) \log p(\text{CaF}_2) + \ldots
$$

+ p(\text{ZrSi}_2) \log p(\text{ZrSi}_2)\}
The entropy after the first partitioning, $H_a$, is thus calculated by:

$$H_a = \{P_1(H_{\text{left node}}) + P_2(H_{\text{right node}})\}$$

$$= \{(221/840)(3.4954) + (619/840)(3.893)\}$$

$$= 3.7884 \text{ bits}$$

The information gain, $IG = H_R - H_a = 4.3751 - 3.7884 = 0.5867 \text{ bits}$

The size of the classification tree should be optimized, so as not to grow excessively large without better efficiency. To avoid overfitting, the pruning of the tree is followed through the cross-validation until the performance of the classification is not improved [4.15]. We tested our model by a random-sampling-based cross-validation procedure; i.e. first a part of the data is randomly selected and set aside for the test, and a classification tree is formed using the remainder called a training data set. Then using the test data, we check out whether the predictive model, i.e. classification tree, suggests the correct structure type. This procedure is repeated in turn using the whole dataset. The validation process is carried out at various tree levels, and the partitioning is stopped at the level of which the prediction error rate is the lowest. Once the prediction criteria are constructed, the structure of a new compound can be estimated by using the structure-prediction model. The entropy change at each step of partitioning is calculated and summed up for both the respective APPs and crystal-structure types in order to quantitatively measure the contribution of the respective APPs on the partitioning process. Crystal-chemistry rules on the regularities in the formation of stable crystal structure of compounds are generated by the classification rules obtained from a recursive process of the partitioning.
4.3. Results and discussion

4.3.1. Assessment of atomic, physical parameters by information entropy

The selection of appropriate atomic, physical parameters (APPs) is essential to construct a good classification model. In this study, we employed seven APP parameters suggested by Mooser-Pearson, Miedema, and Villars [4.6-4.8]. By choosing the parameters well-known as the factors responsible for the structural stability, we attempted to show how 1) to evaluate the relative contribution of the APPs as the influencing factors for crystal structures, 2) to establish the predictive classification model, and 3) to extract valid crystal-chemistry rules from the multivariate data. First three parameters (VEC, $\Delta X_{MB}^{MB}$, $\Delta R_{Zs+p}^{Zs+p}$) were introduced by Villars for the classification of stoichiometric binary compounds, i.e. AB, AB$_2$, AB$_3$, A$_3$B$_5$, etc. [4.10]. The next two parameters ($n_{av}$, $\Delta X_{Pauling}^{Pauling}$) were selected by Mooser-Pearson to describe the structural stability of valence compounds [4.6]. The last two parameters ($\Delta \Phi^*$, $\Delta n_{ws}^{3/3}$) were designed by Miedema for the estimation of the heat of formation of intermetallic compounds [4.7]. The brief description of each parameter was summarized in Table 4.2.

So far various types of parameters have been developed in order to search the regularities regarding the structural stability of compounds. In the systematics of Villars [4.5], several different combinations of APP parameters, as each being involved in one of the five factor groups, can be selected as the model parameters. The difference of Zunger’s pseudopotential radii ($R_{Zs+p}^{Zs+p}$) in the size-factor group and valence-electron concentration (VEC) in the valence-electron factor group are the example. Similarly, average principal quantum number ($n_{av}$) is included in the
atomic-number factor group.

Table 4.2. Atomic parameters for the classification of crystal structure data [4.6-4.8].

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>VEC</td>
<td>Average number of valence electrons per atom</td>
<td></td>
</tr>
<tr>
<td>$\Delta X^{MB}$</td>
<td>Weighted difference of Martynov-Batsanov electronegativities</td>
<td>Villars</td>
</tr>
<tr>
<td>$\Delta R_{zs+p}$</td>
<td>Weighted difference of Zunger’s pseudopotential radii sum</td>
<td></td>
</tr>
<tr>
<td>$n_{av}$</td>
<td>Average principal quantum number</td>
<td>Mooser &amp; Pearson</td>
</tr>
<tr>
<td>$\Delta X^{Pauling}$</td>
<td>Pauling electronegativity difference</td>
<td></td>
</tr>
<tr>
<td>$\Delta \Phi^*$</td>
<td>Chemical potential difference for electronic charges</td>
<td>Miedema</td>
</tr>
<tr>
<td>$\Delta n_{WS}^{1/3}$</td>
<td>Electron density difference in Wigner-Seitz atomic cell</td>
<td></td>
</tr>
</tbody>
</table>

Four out of seven parameters we used herein, i.e. two electronegativity scales (after Pauling, Martynov-Batsanov), chemical potential (after Miedema), and electron density in a Wigner-Seitz cell (after Miedema) correspond to the electrochemical-factor group. We may choose ionic, covalent, or metallic radii as the size factor instead of Zunger’s pseudopotential radii ($R_{zs+p}$) in other datasets, considering the bonding states of the system. How to recognize the relative significance of a rather broad combination of the parameters to select the best one? In many cases, the parameters are interrelated and thus not completely independent on each other. Furthermore, since some of the parameters adopted have been employed to describe solely specific type of compound groups, a set of parameters which successfully classifies one-type of compounds might fail when applying for other kinds. For example, according to J. K. Burdett et. al. [4.9], the radii ratio of sigma- and pi-
bonding, $R_\sigma/R_\pi$, is not a good parameter for the classification of $AB_2$ type compounds in spite of its success in the discrimination of $AB$ octet compounds, instead the ratio, $r_\sigma^A/r_\sigma^B$, is a good parameter for the $AB_2$ compounds. Wherein, $R_\sigma = |r_\sigma^A-r_\sigma^B|$ and $R_\pi = r_\pi^A+r_\pi^B$, and $r_\sigma$ and $r_\pi$ are the summation and difference of the pseudopotential radii of $(s+p)$ electrons, respectively. Thus, it is crucial to select effective APPs among the numerous parameters that are expected to influence on the formation of stable crystal structures.

In this regard, the application of a systematic evaluation algorithm for the selection of APPs is necessary for a large number of parameters that have already been developed. Information-entropy-based classification provides the criterion for the selection of effective parameters. According to this method, multiple parameters are variably used for the classification of crystal structure data. In brief, different sets of attributes play the role for the classification of crystal structure data. The magnitude of the influence of respective APP expressions on the occurrence of a specific structure type can be measured by the relative contribution of the APPs to the classification of structure data. Figure 4.2 shows the evaluation result of the attributes in which the contribution is quantitatively measured in terms of the change of information entropy for the classification of structure data. As previously mentioned, the seven parameters used are involved in one of the four group categories after Villars; that is, valence-electron factor (VE), size-factor (SZ), electrochemical factor (EC), and atomic-number factor (AN) groups. Four parameters ($\Delta X^{\text{Pauling}}, \Delta X^{\text{MB}}, \Delta \Phi^*, \Delta n_{wz}^{1/3}$) correspond to the EC factor group, whereas for the SZ ($\Delta R^{s+p}$), VE (VEC), AN ($n_{av}$) factor groups one parameter is tested, respectively. Whereas the contribution of VE factor is predominant, the effects of SZ factor and EC factor compete each other. However, one thing which should be noted
is that the determination of a certain structure type is controlled by the combined effects of the multiple factors, not by a single factor only, even when we call size-factor compounds, electron compounds, or valence compounds, etc. for specific compound groups. The seven APPs are not equally significant for the determination of different crystal structures. For example, among the three types of electronegativity scale tested in this study, i.e., the scales suggested by Pauling ($\Delta X^{Pauling}$), Martynov-Batsanov ($\Delta X^{MB}$), and Miedema ($\Delta \Phi^*$), the contribution of Pauling's electronegativity is dominant; note that the weighted average values were used for $\Delta X^{MB}$, whereas non-weighted difference was used for $\Delta X^{Pauling}$. According to Villars, non-weighted $\Delta X^{MB}$ is more effective than the weighted expression for the structure mapping of $AB_2$-type compounds [4.10]. Between the average valence electron numbers (VEC in Villars’ model) and electron density in a Wagner-Seitz cell ($\Delta n_{ws1/3}$ in Miedema’s model), the former plays the major role in the classification. It is also observed that the average principal quantum number ($n_{av}$), one of the seven parameters, is rarely used for the classification. Whereas the structure data we used were mainly of intermetallic compounds, $n_{av}$ in the Mooser-Pearson model was used for the separation of valence compounds with ionic- and covalent-type bonding. Compared with the parameters of other two models, the reduction of the information entropy due to the Villars’ model parameters is dominant in general. As previously mentioned, however, it is mainly due to the fact that the parameters from three different models were originally devised as the structure-map coordinates which classify the different types of compounds.
Figure 4.2. Evaluation of atomic and physical parameters from three structure-map models. Each parameter contributes to the reduction of the entropy with different extent. Atomic parameters used in this paper can be categorized into four factors (groups); valence-electron factor (blue), size factor (red), electrochemical factor (green), and atomic-number factor (yellow). Average principal quantum number ($n_{av}$) does not contribute for the classification of data. Based on this result, only three factors will be discussed below in detail.

4.3.2. Classification rule mining of crystal chemistry

While the relative contribution of APPs is assessed in terms of the information-entropy change, we could deduce a number of new rules of crystal chemistry with respect to the relationship of structure-governing APPs to the crystal structure of compounds. As already shown in Figure 4.2, the contribution of the valence-electron factor (VE) is generally predominant on the discrimination of most of the crystal structure types. Note that the high contribution of a specific APP, i.e. a large reduction of information entropy due to a particular APP, implies that a specific structure-type domain in the chemical space can be well discriminated by the
parameter from other structure domains. Therefore, the partitioning corresponds to the process of identifying relevant attribute dimensions for each class of structure types.

As compared the size factor (i.e., $\Delta R_{\text{size}}^2$) and electrochemical factors (i.e., $\Delta X^{\text{MB}}$, $\Delta X^{\text{Pauling}}$, $\Delta \Phi^*$, and $\Delta n_{\text{ws}}^{1/3}$), it is observed that different APPs play the role as the major contributing factor depending on the structure type of the compounds. For example, Lave phases, the well-known size-factor compounds [4.15], i.e., MgCu$_2$ and MgZn$_2$ structures show the large magnitude of entropy change originated from the size-factor term whereas the contribution of electrochemical factors is much lower than that of the size factor (Figure 4.3(a)). This observation is in a good agreement with the results from other references [4.15, 4.16]. The competing structure types of MgCu$_2$ and MgZn$_2$, i.e. AlB$_2$ and CuAl$_2$ structure types also represent the higher contribution of size factor than that of the electrochemical factors. However, in the case of the MoSi$_2$, another competing structure type of Laves phase, and KHg$_2$, an AlB$_2$-related structure, the difference is not apparent. Another example, Co$_2$Si-b, PbCl$_2$ and Ni$_2$In shows rather different trends (Figure 4.3(b)) particularly with related to the contribution of VE term. In these three structure types, the contribution of the average valence-electron number is rather diminished and not so critical. Instead, the contribution of size-factor (for Co$_2$Si-b and Ni$_2$In) and electrochemical factors (for PbCl$_2$) are significant as the dominating term, respectively. Two structure types, Ni$_2$In and Co$_2$Si-b show similar pattern of APP contribution; if the unit cell of Ni$_2$In structure, a superstructure of AlB$_2$ [4.17], is distorted, then it becomes Co$_2$Si-b structure. A structure type which includes many Zintl-phase compounds, CaF$_2$, shows rather higher contribution of electrochemical factor term, and CaC$_2$ structure also shows high contribution of electrochemical
factor. For example, two Zintl compounds, Mg₂Ge and Mg₂Sn have anti-CaF₂-type structure, and MgEr₂ and MgTb₂ have MoSi₂-type structure. The information entropy measure for these structure types shows that electrochemical factor In case of the structure types MoSi₂ and CaC₂, the distorted form of MoSi₂, the contribution of size and electrochemical factor shows dissimilar pattern. It is thought that because the distortion of the structure is in fact large in magnitude enough to change the coordination, even though those are isopointal structures [4.18]. CaIn₂ type, a derivative structure of the AlB₂ structure, is also showed similar pattern with AlB₂ as a size-factor group.
Figure 4.3. The contribution of atomic-property parameters to the respective structure types. (a) Laves phases (MgCu$_2$ and MgZn$_2$ type compounds) and their competing phases (CuAl$_2$ and MoSi$_2$ structure types), (b) Competing structure types. The ‘anti-CaF$_2$’ and ‘anti-PbCl$_2$’ type compounds were also included in CaF$_2$ and PbCl$_2$ structure-type data, respectively. Where, ‘anti-CaF$_2$’ and ‘anti-PbCl$_2$’, imply that the position of the elements corresponding to F or Cl sites are occupied by the group 2 elements (i.e., Mg, Ca, Sr, and Ba) and that of the elements corresponding to Ca or Pb are occupied by group 14 elements (i.e., Si, Ge, Sn, and Pb) [4.2].
Figure 4.4. Information-entropy diagram: the relative contributions of structure-governing factors for \( \text{AB}_2 \) compounds. Each point (▲) indicates the crystal-structure type. For most of structure types, the contribution of valence-electron factor dominates, and the size factor and electrochemical factor competes each other.

The synoptic mapping of the information on the structure formation in Figure 4.4 reflects the relative contribution in varying degrees of three structural factor groups,
i.e. size, valence-electron, and electrochemical factor† for the classification of 34 structure types of AB₂ compounds. Based on the result of Figure 4.2, the effect of atomic-number factor (average principal quantum number) was not taken into account in this plot. According to Villars [4.8], only three factors among the five groups can be effectively used for the description. As aforementioned, most of the structure types are closely placed to near the corner of valence-electron factor. However, structure types such as La₂Sb, Cu₂Sb, and Co₂P show very low contribution of valence-electron factor. The results are compiled in Table 4.3.

<table>
<thead>
<tr>
<th>Governing factor</th>
<th>Structure types</th>
</tr>
</thead>
<tbody>
<tr>
<td>VE factor</td>
<td></td>
</tr>
<tr>
<td>SZ-factor</td>
<td>AlB₂, CaIn₂, CuAl₂, HfGa₂, MgCu₂, MgZn₂, MoPt₂, OsGe₂, ZrGa₂</td>
</tr>
<tr>
<td>EC-factor</td>
<td>CaC₂, CaF₂, Cd₂Ce, CoSb₂, CrSi₂, FeS₂, Hg₂U, HoSb₂, KHg₂, LaSb₂, Mg₂Cu, MoSi₂, NdAs₂, ThSi₂, TiSi₂, ZrSi₂</td>
</tr>
<tr>
<td>SZ factor</td>
<td>Co₂P, Co₂Si-b, Cu₂Sb, CuZr₂, Fe₂P, Ni₂In, Ti₂Ni</td>
</tr>
<tr>
<td>EC factor</td>
<td>La₂Sb, PbCl₂</td>
</tr>
</tbody>
</table>

† In case of the electrochemical factor, the amount of entropy change due to the classification by the four APPs (ΔX_{Pauling}, ΔX_{MB}, ΔΦ*, Δn_{ws}^{1/3}) was summed up for the respective structure types.
Figure 4.5. Information-entropy diagram: (a) some representative structure types of Laves phases (SZ-factor-dominated compounds) and Zintl phases (EC-factor-dominated compounds) are labeled on the diagram. The contribution of VE (valence-electron) factor can be compared with other two factors; (b) AlB$_2$ derivative structures are labeled on the diagram according to the contribution of structure-governing factors. Whereas structure family tree shows the symmetrical relations, this diagram reflects crystal-chemical relations. In case of KHg$_2$ structure, SZ-factor and EC factor compete against each other (the contribution of EC factor is rather higher than that of SZ factor).
In this chapter, we have demonstrated that how to assess the factors and principles which determine the crystal structure of compounds by means of the classification of crystal structure data. The multivariate classification algorithm equipped with quantitative evaluation criteria for the attributes facilitated both the test of the attributes used and the search of embedded rules from data, extracting the information that human inference may fail to recognize. The question of what kinds of atomic, physical parameters (APPs) should be used as the attributes is of critical importance, since it ultimately determines what information (principles) we can finally see. Although only seven APP expressions categorized into four different groups were used in this study, other kinds of parameter sets are equally applicable for the classification in a same manner. In effect, there is no single panacea for the effective selection of the factors. After collecting all possible APP expressions that are suspected to influence the formation of structures, all possible combinations of the parameters need to be tested to ensure whether we are free from missing the “best solution” by missing the consideration of any “best” combination of APP expressions. We also showed that the crystal-chemical rules achieved by the machine-learning algorithm can provide the additional information with respect to the mechanism of the structure formation. Conclusively, this rule mining scheme can play the role as a methodology which complements the purely human-based learning, in that not only does it accelerate the learning process but further it may disclose some hidden rules beyond human intuition and inference.
References


5.1. Integration of recursive partitioning and first principles calculations: prediction of the stable crystal structures

In spite of the advancement of quantum theory and computational foundations, the prediction of the stable structure of any new material through quantum-mechanical calculations of the total energy of the system is still not an easy task. There are two major reasons for it. First, the energy differences among dissimilar crystal structures are too small as compared to the total energy of a system (i.e. of the order of one hundredth) and the energy of an isolated atom is much higher than the cohesive energy in the solid state. Second, in spite of the modern super-computing power, there are too many local (energy) minimum states to be excluded in the energy landscape and we might need infinite number of computations to test all possible atomic arrangements. Here, the data-driven approach such as the method introduced in this thesis comes in. Once the classification of the given data is implemented, the rules achieved from the classification can be used as the guideline for the estimation of the stable structures of compound materials. Since diverse crystal structures form overlapping regions in the (atomic) parameter space, the classification mapping provides - most of the cases - a list of some possible crystal structures for a compound rather than gives a single most possible structure as the answer. The partial overlapping of crystal-structure domains (here, a structure domain is defined as a well-defined region for a structure type) or some exceptional
compounds sitting in different structure domains in the atomic-parameter space are mainly due to three reasons; (i) the two or more than two structures can be stable, (ii) the structural stability of some compounds are not only determined by the atomic parameters, and (iii) the overlapping structures are related to each other in the crystallographic aspect. The value of this type of approach is to (i) make groups with the compounds based on some similarities of structure types, and (ii) to suggest which crystal structures should be calculated by first principles methods. How the combination of these two methods can be used for the prediction of crystal structures will be described below.

**Figure 5.1.** The integration of two principles for the estimation of the stable crystal structure; data mining provides the list of possible stable structures, reducing the computation load of ab initio approach. Then ab initio calculations find a crystal structure with the lowest formation energy among the candidates suggested by data mining.

### 5.1.1. Energy calculation by density functional theory

Ab-initio calculations of the electronic structure were performed within the framework of the density functional theory (DFT) by using the projector augmented-wave method (PAW) as implemented in VASP [3.28]. We used PAW potentials [3.29, 3.30] derived within the generalized gradient approximation (GGA) description of
the electronic exchange-correlation energy \[3.31\]. All structures were fully relaxed until the Hellman-Feynman forces acting on each ion became less than \(10^{-3}\) eV/Å. To ensure accurate results during the structure optimization procedure, Kohn–Sham orbitals were expanded in a plane wave basis up to an energy cutoff \(3/2\) larger than the default energy cutoff provided by PAW potentials. We used Monkhorst-Pack \[3.32\] scheme to generate an automatic k-mesh sampling of the Brillouin zone and the integration in reciprocal space was carried out by using the Methfessel–Paxton \[3.33\] smearing during the relaxation and the linear tetrahedron method with Blöchl corrections \[3.34\] for the relaxed structures. For all structures the convergence within \(10^{-3}\)eV/ion of the total energy with respect to the number of k-points was achieved.

5.1.2. Prediction of stable crystal structures

Figure 5.2 illustrates how to find the most stable structure of a hypothetical compound, AuBe\(_2\), by combining our classification-tree model and first principles calculations. First, the parameters affecting crystal structures (i.e. APPs) of AuBe\(_2\) are calculated from the atomic properties of the relevant constituent elements. Then the APPs of the test compound are compared with the criteria shown in the classification tree, which is tracked from the root node (“start” node in Figure 3.9) to the leaves. At the end of a leaf, a list of the crystal structure types, AlB\(_2\), CaF\(_2\), MgCu\(_2\), OsGe\(_2\), and PbCl\(_2\), is suggested as the possible stable structures of AuBe\(_2\). That is, one of these five structures is the answer we are looking for. This prediction procedure is completed through the confirmation by ab initio total energy calculations. The results of the calculation suggest that the structure type MgCu\(_2\), which has the lowest energy, is the most stable structure for AuBe\(_2\). In the same way, the crystal
structure of some compounds – experimentally known but the structure types are not yet ascertained – has been predicted and the results were summarized in Table 5.1. Most of the cases, only two to seven structures are suggested as the candidates of possible stable structures, and density-functional theoretical (DFT) calculations are implemented only for the atomic configurations suggested by the if-then rules of the classification tree.

Figure 5.2. Prediction of the crystal structure of AuBe₂. As searching through a route which meets the values of the physical parameters of a test compound, one to ten structure types are nominated as the possible crystal structure candidates. (1) The parameters calculated from atomic and physical properties of the constituents of AuBe₂ are used as the input variables for the structure prediction; (2) The classification tree suggests the possible structure-type candidates (the pathway along the red arrows); (3) A structure type with the lowest total energy, i.e. MgCu₂, is confirmed as the most stable crystal structure by first principles calculations.
Table 5.1. Prediction of the crystal structure (prototype) of AB\textsubscript{2} type compounds.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Structures estimated by classification tree</th>
<th>Structure predicted by DFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlRu\textsubscript{2}</td>
<td>CuZr\textsubscript{2}, Cu\textsubscript{2}Sb</td>
<td>CuZr\textsubscript{2}</td>
</tr>
<tr>
<td>FeMn\textsubscript{2}</td>
<td>CuZr\textsubscript{2}, CuAl\textsubscript{2}</td>
<td>CuZr\textsubscript{2}</td>
</tr>
<tr>
<td>FeB\textsubscript{2}</td>
<td>PbCl\textsubscript{2}, CaF\textsubscript{2}, HoSb\textsubscript{2}, LaSb\textsubscript{2}, NdAs\textsubscript{2}</td>
<td>PbCl\textsubscript{2}</td>
</tr>
<tr>
<td>Co\textsubscript{2}Na</td>
<td>MgZn\textsubscript{2}, FeS\textsubscript{2}, MgCu\textsubscript{2}</td>
<td>MgZn\textsubscript{2}</td>
</tr>
<tr>
<td>AuCa\textsubscript{2}</td>
<td>PbCl\textsubscript{2}, Co\textsubscript{2}P, CuAl\textsubscript{2}, CuZr\textsubscript{2}, Hg\textsubscript{2}U, MgCu\textsubscript{2}, MoSi\textsubscript{2}</td>
<td>PbCl\textsubscript{2}</td>
</tr>
<tr>
<td>Au\textsubscript{2}Cr</td>
<td>PbCl\textsubscript{2}, Co\textsubscript{2}P, CuAl\textsubscript{2}, CuZr\textsubscript{2}, La\textsubscript{2}Sb, MoSi\textsubscript{2}, Ti\textsubscript{2}Ni</td>
<td>PbCl\textsubscript{2}</td>
</tr>
<tr>
<td>CsGa\textsubscript{2}</td>
<td>KHg\textsubscript{2}, AlB\textsubscript{2}, CaC\textsubscript{2}, CaF\textsubscript{2}, MgCu\textsubscript{2}, ThSi\textsubscript{2}</td>
<td>KHg\textsubscript{2}</td>
</tr>
<tr>
<td>GdPd\textsubscript{5}</td>
<td>Ni\textsubscript{3}In, AlB\textsubscript{2}, Co\textsubscript{2}Si-b, MgCu\textsubscript{2}, MgZn\textsubscript{2}</td>
<td>Ni\textsubscript{3}In</td>
</tr>
<tr>
<td>RuGe\textsubscript{2}</td>
<td>OsGe\textsubscript{2}, AlB\textsubscript{2}, CaF\textsubscript{2}, MgCu\textsubscript{2}, PbCl\textsubscript{2}</td>
<td>OsGe\textsubscript{2}</td>
</tr>
</tbody>
</table>

5.2. Phase-transition paths of metal-hydride alloys

At present, most of the hydrogen storage alloys are based on AB\textsubscript{m} type (m=0.5, 1, 2, 3, and 5) intermetallic compounds as the host material; for instance, the host compounds have AB (e.g. TiFe), AB\textsubscript{2} (e.g. TiCr\textsubscript{2}, ZrCr\textsubscript{2}, ZrFe\textsubscript{2}, ZrMn\textsubscript{2}), A\textsubscript{2}B (e.g. Mg\textsubscript{2}Ni, Ti\textsubscript{2}Ni), AB\textsubscript{5} (e.g. LaNi\textsubscript{5}, CaNi\textsubscript{5}) type composition formulae. During the hydride formation, i.e. AB\textsubscript{m} → AB\textsubscript{m}H\textsubscript{n}, the lattice structure of the host material is whether expanded (the swelling of the lattice) or transformed to other crystal structures by the hydrogen uptake. Since the practical usage of the hydrogen-storage alloys involves the cycling of hydrogen uptake/release, it is expected that the repeated structural transitions of the materials would be accompanied with the hydrogenation ↔ dehydrogenation process. Developing new hydrogen storage materials, therefore it is necessary to investigate materials systems considering the structural outline.
At the Chapter 2, the similarity and relationships of the crystal structures of AB$_2$ binary compounds have been investigated using crystallographic group theory. In this section, we apply the structure-family tree (group-subgroup relations) in order to investigate structural change of intermetallic compounds during the hydrogenation and provide the guideline for the rational design of intermetallic compound hydrides. The design rules derived from the group-theoretical interpretation would ultimately offer the possible new directions for searching of new hydrogen-storage alloys.

Hydrogen-storage materials should meet two objectives; that is, i) fast rates of absorption/desorption processes, and ii) high hydrogen content. The capacity of the hydride materials is presented in the stoichiometric ratio of the hydrogen to metal (H/M) and gravimetric capacity (wt.% = H/(H+M)×100). Most of the efforts to find the design rules for new hydrogen storage alloys in the crystal-structural aspect have not been satisfactory. Two empirical criteria suggested by D. G. Westlake [5.3], that is, the minimum vacancy size of 0.40 Å and minimum distance between hydrogen atoms of 2.1 Å in the host matrix, were not enough for the structural design of new hydrides for the practical purpose. Also, the phase stability of hydrogen-storage alloys should be optimized at the intermediate level, implying that both the hydrides either too stable or too unstable are not so suitable because most of the practical applications require the hydrogen absorption/desorption cycling under mild conditions of temperature and pressure.

As briefly mentioned, insertion of hydrogen atoms into the intermetallic compounds causes the change of the crystal structure of host materials in their atomic level (Figure 5.3). The structural changes during the repeated sorption process are
therefore highly related to the sorption kinetics as well as the mechanical stability. The structural transformation is proceeded through one of the two pathways; reconstructive and displacive transformations. In reconstructive transformation, the chemical bonds are broken and atoms are rearranged forming a new structure. It implies that the structural change would require high activation energy. On the contrary, displacive transformation does not involve the breaking of bonds and thus usually is accompanied with a rapid kinetics of the displacement of atoms under moderate conditions. In terms of symmetry relations, reconstructive transformation does not involve crystallographic group-subgroup relations between the structures, but displacive transformation does. A pair of intermetallic compounds and the corresponding hydrides connected through the phase-transition paths with group-subgroup relations are preferable for the practical application as a hydrogen-storage material which requires the absorption-desorption cycling.

**Figure 5.3.** A schematic description of the structural changes according to the hydrogenation-dehydrogenation cycling process (this figure was modified from the ref. 5.4). Green-color dots denote hydrogen atoms. As the structural transition 1 ↔ 2 have the group-subgroup relations, the reversible displacive phase transition is undergone during the hydrogen absorption/desorption process.
In this study, the change of crystallographic structures between intermetallic compounds and the corresponding hydrides is investigated by means of the group-subgroup relations. Although the rules based on these symmetry relations cannot suggest energetically more probable pathways of the structural change, they provide the guideline for finding more favorable structural transition paths which meet the practical cycling conditions.

A magnesium-based intermetallic hydride, Mg$_2$FeH$_6$, has been attracted as a hydrogen-storage material due to its high hydrogen content (~5.6 wt.% hydrogen). However, the host Mg$_2$Fe does not form a stable phase without the inclusion of hydrogen and this fact becomes a critical drawback which leads the slow kinetics of the reversible process of hydrogen uptake/release. Here, the classification tree (Figure 3.7) and structure family tree (Figure 2.9) will be used as the guideline to proceed towards the next step of Mg-Fe-H systems.

Using the classification tree of AB$_2$-type compounds of Figure 3.7, the possible crystal structure of Mg$_2$Fe can be estimated as one of the following seven structure types; that is, CaF$_2$ (Fm$ar{3}$m, 225), Co$_2$P (Pnma, 62), Cu$_2$Sb (P4/nmm, 129), CuAl$_2$ (I4/mcm, 140), CuZr$_2$ (I4/mmm, 139), Ni$_2$In (P6$_3$/mmc, 194), and Ti$_2$Ni (Fd$ar{3}$m, 227). The pathway of the classification tree which provides the possible crystal structures of hypothetical Mg2Fe system is shown in Figure 5.4. Then, the next step is to find crystal structure types which have the group-subgroup relations with the CaF$_2$ structure.
Figure 5.4. A part of the classification tree for AB\textsubscript{2} compounds. The if-then rules for the prediction of the crystal structure of Mg\textsubscript{2}Fe are shown based on the calculated APP values of the hypothetical compound, Mg\textsubscript{2}Fe. As the possible structure, seven crystal structure types are suggested from the classification tree. Among them, favorable structures for the hydrogen uptake/release process are selected by using group-subgroup relations shown from the structure family tree.

Conclusively, according to the structure family tree of the group-subgroup relations for the structure type CaF\textsubscript{2} (of the hydride, Mg\textsubscript{2}FeH\textsubscript{6}) in Figure 2.9, the most preferable structure types of Mg-Fe-based host materials for the reversible hydrogenation/dehydrogenation process correspond to CaF\textsubscript{2} (Fm\textbar 3m, 225) or CuZr\textsubscript{2} (I\textbar 4/mmm, 139) among the seven candidate structures from the classification tree. The following is the part of the structure family tree of AB\textsubscript{2} type structures, which shows four structure types related with CaF\textsubscript{2} structure in terms of group-subgroup relations. Therefore, the intersection of the two sets, \{CaF\textsubscript{2}, Co\textsubscript{1}P, Cu\textsubscript{2}Sb, CuAl\textsubscript{2}, CuZr\textsubscript{2}, Ni\textsubscript{2}In, and Ti\textsubscript{2}Ni\} ∩ \{CaC\textsubscript{2}, CuZr\textsubscript{2}, La\textsubscript{2}Sb, and MoSi\textsubscript{2}\} = \{CuZr\textsubscript{2}, (and CaF\textsubscript{2} itself)\}
The symmetrical relations for a possible structural transition during the hydrogen “absorption ↔ desorption” process are shown in the Figure 5.5, using CaF$_2$ (Fm$ar{3}$m) ↔ CuZr$_2$ (I4/mmm) transformation.

**Figure 5.5.** The symmetrical relations between the crystal structure of Mg$_2$FeH$_6$ and a hypothetical structure of the corresponding host Mg$_2$Fe compound; during the hydrogen uptake/release cycle, the structural change undergoes through either no transition or group-subgroup transition.
References


CHAPTER 6
CONCLUSIONS: CONTRIBUTIONS OF THIS STUDY

Data mining has been developed for a long time as an instrumental methodology in diverse fields. However, its serious advent in the materials research has rather recently been brought in. To settle down the data mining as a main stream of computational materials science, the comprehensive understanding on the connections between data mining and materials science would be demanded along with more examples of its successful applications. In this regard, the major contributions of my research which have been shown in this thesis are as follows.

First, in this study, data-driven approaches have been applied to the crystal chemistry study as a pragmatically effective tool of computational research. Integrated with quantum-mechanical calculations, the rules extracted from the crystal structure data have been built as a model for the prediction of crystal structures of inorganic crystalline compounds. The guideline provided by this experimental-data-driven model has extensively reduced the workload imposed upon ab initio calculations for the search of stable crystal structures of novel materials.

Second, for the first time, the formation mechanism of different crystal structures of inorganic compounds has been investigated by using information-theoretic classification of materials data. The relative contribution of atomic, physical
parameters (APPs) of chemical constituents upon the structural stability of the corresponding compound materials has been quantitatively measured in terms of information entropy.

Third, the similarity (or dissimilarity) and connections among the crystal structures of \( \text{AB}_2 \) compounds were investigated by applying crystallographic-group theory. The network diagram which represents the symmetry (supergroup-subgroup) relations could be used for the materials design.

Finally, this study provides a representative example, which is pragmatically valuable, of the application of data-driven approach for materials design. The methods shown in this thesis can be applied to a variety of material behaviors including structure-property relations.
A1.1. Introduction

Recursive Partitioning divides the feature space into a set of rectangles based on the relationship between the attributes and the predefined classes. That is, attribute space is subdivided into smaller parts called *bins* in a way that similar members group together and those dissimilar are divided into separate bins. These splits recursively performed are represented as a tree structure. Through this partitioning-based classification, one can achieve a series of rules, i.e. *classification rules*, from data for the prediction or learning, which enables one to readily understand the structure, characteristics, and the relationships of correlation and causation. The classification rules achieved from training datasets determine the class of any new system based upon its values of the attributes. It is not only conceptually simple for the interpretation of the result, but also is extremely a powerful technique for the handling of large data sets.

The recursive partitioning method uses different statistical measures depending upon whether the type of variables is categorical or continuous. When the variables are categorical, the partitioning method is called “Classification Tree” since it is based on predicting classes in each of the sub-groups. Whereas, when the variables are continuous, it is called “Regression Tree” since we are trying to predict the numerical value for each of the sub-groups (Figure A1.1).
Figure A1.1. Recursive partitioning: the graphical representation of the results of recursive partitioning can be categorized into two kinds, i.e. classification and regression trees, depending on whether the data analyzed consist of discrete-valued (categorical or numerical) or continuous-valued (numerical) outputs.

A1.2. Classification tree

In the case of a classification tree, we are trying to identify subsets that predominantly belong to one particular class. In a node $m$, let $p_{mk}$ represents the proportion of class $k$ observations in a node $m$. We classify the observations in node $m$ to the majority class $k$ in a node $m$, given by

$$x_i(m) = \underset{x_i}{\text{argmax}}(p_m(x_i))$$

Where, $x_i(m)$ is a major class (i.e., structure type) in a node $m$ and $p_m(x_i)$ is a probability of the class $x_i$ in the node $m$, and “argmax” denotes the “argument of the maximum”; namely, as the classification tree grows, the homogeneity of each node is maximized. The general idea of recursive partitioning is to reduce the feature space into subsets of low impurity. The partitioning and the corresponding output values for each subset depend on the measure of impurity. Different Measures of impurity
is used depending upon whether it is a classification or a regression problem.

**A1.2.1. Solving for the best partitions: Impurity function**

In general, impurity function, $H$, has a concave shape and can be defined as:

$$ H(t) = Q_m(p(c | t)) $$

Where, the function $Q_m$ has the properties (1) $Q_m \geq 0$ and (2) for any $p \in (0,1)$, $Q_m(p) = Q_m(1-p)$ and $Q_m(0) = Q_m(1) < Q_m(p)$.

There is no single justification for the use of specific function. However, the function to measure the “purer” descendant nodes than the data in the ascendant set should have the following properties.

(1) When defining the node proportions $p(c_i | t)$, where $c_i = c_1, c_2, \ldots, c_n$, to be the proportion of the cases $X \in t$ belong to class $c_i$ so that

$$ p(1 | t) + p(2 | t) + \ldots + p(n | t) = 1 $$

(2) When defining a measure $H(t)$ of the impurity of $t$ as a nonnegative function $Q_m$ of the $p(1 | t), p(2 | t), \ldots, p(n | t)$ such that

$$ Q_m(1/n, 1/n, \ldots, 1/n) = \text{Maximum} $$

$$ Q_m(1, 0, 0, \ldots, nth 0) = 0, Q_m(0, 1, 0, \ldots, nth 0) = 0, \ldots, Q_m(0, 0, 0, \ldots, 1) = 0 $$

That is, the impurity of a node should be largest when all classes are equally mixed together and smallest when each node includes only one class.
The functions commonly chosen include (in two-class problem):

1. \( Q_m(p) = \min(p, 1-p) \)  
   \[ \text{Bayes error} \]

2. \( Q_m(p) = -p \log(p) - (1-p) \log(1-p) \)  
   \[ \text{Entropy function} \]

3. \( Q_m(p) = p(1-p) \)  
   \[ \text{Gini index} \]

These functions can be illustrated by the following Figure A1.2.

![Figure A1.2. Impurity functions [A1.3].](image)

The general form of the measures for the impurity of the node \( m \), \( Q_m(T) \) are:

**Bayes error:**  
\[ Q_m(p_{mk}) = 1 - p_{mk}(m) \]

**Gini index:**  
\[ Q_m(P_{mk}) = \sum_{k \neq k'} p_{mk} p_{mk'} = \sum_{k=1}^{K} p_{mk} (1 - p_{mk}) \]

**Information entropy:**  
\[ Q_m(p_{mk}) = -\sum_{k=1}^{K} p_{mk} \log_2 p_{mk} \]

Information entropy and Gini index are more sensitive to changes in the node probabilities than the Bayes' misclassification error and hence Gini index or information entropy are preferred over misclassification error for growing the classification trees. Gini index is also known that the size of the tree branches is often developed in unbalanced way [A1.3]. Note that both of these functions have their minimum, 0, when \( p=1 \) for some \( t \) while the other \( p_j \), \( j \neq t \), are zeros, and that both of them have their maximum when all \( p_i \) are equal.
A1.3. Algorithmic framework

Building a classification tree follows the general procedure of the partitioning as shown below.

![Pseudo Code]

1. Start with Root node.
2. Split with the attributes which lead the maximum information gain.
3. Fully grow the tree by splitting best recursively.
4. Prune worst nodes using impurity measures.
5. Choose the optimal tree structure which makes the best predictability.

A1.3.1. Recursive partitioning of a two-class problem

The following Figures A1.3 show the distribution of data entries of AB$_2$ type compounds with CuAl$_2$ and KHg$_2$ structure types according to two different APPs. In the case of the top figure which uses the electronegativity difference, $\Delta X_{MB}$, as the X-coordinate (for the separation of two structure types), it is clearly shown that the distributions of both crystal structures are overlapped in most of the range of the electronegativity difference, implying that this parameter is not suitable for the differentiation of these two structures. The bottom figure shows similarly the distributions of the two structures according to the atomic size (i.e. pseudopotential radius) difference, $\Delta R_{sp}$; in this case, the distributions are rather well-separated although there is a partial overlapped region. Between the two parameters, $\Delta X_{MB}$ and $\Delta R_{sp}$, thus, $\Delta R_{sp}$ is a better classifier. To find the splitting value of classifier (i.e. $\Delta R_{sp}$), the information gain by the classification can be calculated as follows.

There are the entries of 50 AB$_2$ compounds with CuAl$_2$ structure and 51 compounds with KHg$_2$ structure in the LPF database. Then, the information entropy, $H_{root}$, before
the partition is calculated by:
\[ H_{\text{Root}} = - \sum p(x) \log_2 p(x) \]
\[ = - \{(p(\text{CuAl}_2) \log_2 p(\text{CuAl}_2) + p(\text{KHg}_2) \log_2 p(\text{KHg}_2)) \} \]
\[ = -(50/101) \log_2 (50/101) + (51/101) \log_2 (51/101) \]
\[ = 0.99993 \text{ bits} \]

To find a splitting value which maximize the information gain (IG) due to the first partitioning, the IGs are calculated for the respective \( \Delta R_{s+p} \) values at the range of overlapping region (see the bottom figure of Figure A1.3) of \( \Delta R_{s+p} \) from 0.22664 to 0.54328. \( H_i \) is the entropy after the first partitioning which is calculated as follows:

For instance, let \( \Delta R_{s+p} = 0.22664 \) for the partitioning. At the left side of the splitting line, i.e. \( \Delta R_{s+p} < 0.22664 \), there are 38 compounds with \( \text{CuAl}_2 \) structure. At the right side of the line, i.e. \( \Delta R_{s+p} \geq 0.22664 \), there are 12 \( \text{CuAl}_2 \) and 51 \( \text{KHg}_2 \) structure entries.

Thus, the entropy after the first partitioning, \( H_i \), is:
\[ H_i = -(38/101) \times (38/38) \times \log_2 (38/38) - (12/101) \times (12/63) \times \log_2 (12/63) \]
\[- (51/101) \times (51/63) \times \log_2 (51/63) \]
\[ = 0.17876 \text{ bits} \]

The information gain (IG) is then calculated:
\[ IG = H_{\text{Root}} - H_i \]
\[ = 0.99993 - 0.17876 = 0.82117 \text{ bits} \]

In the same way, all the possible splitting values between 0.22664 and 0.54328 are tested. The results summarized in the following table show the information gain by the partitioning at the \( \Delta R_{s+p} = 0.53328 \) would maximize the information gain, IG=0.90748.
Table A1.1. Determination of the partitioning position by information-entropy calculations; the splitting value of the maximum information gain was highlighted.

<table>
<thead>
<tr>
<th>$\Delta R_{sp}$</th>
<th>$H_{\text{root}}$</th>
<th>$H_1$</th>
<th>Information gain (IG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.22664</td>
<td>0.99993</td>
<td>0.17876</td>
<td>0.82117</td>
</tr>
<tr>
<td>0.23998</td>
<td>0.99993</td>
<td>0.18001</td>
<td>0.81991</td>
</tr>
<tr>
<td>0.24331</td>
<td>0.99993</td>
<td>0.16623</td>
<td>0.83370</td>
</tr>
<tr>
<td>0.27997</td>
<td>0.99993</td>
<td>0.15210</td>
<td>0.84783</td>
</tr>
<tr>
<td>0.28331</td>
<td>0.99993</td>
<td>0.16755</td>
<td>0.83238</td>
</tr>
<tr>
<td>0.28664</td>
<td>0.99993</td>
<td>0.15314</td>
<td>0.84679</td>
</tr>
<tr>
<td>0.30997</td>
<td>0.99993</td>
<td>0.13840</td>
<td>0.86152</td>
</tr>
<tr>
<td>0.36663</td>
<td>0.99993</td>
<td>0.12336</td>
<td>0.87657</td>
</tr>
<tr>
<td>0.37996</td>
<td>0.99993</td>
<td>0.12376</td>
<td>0.87617</td>
</tr>
<tr>
<td>0.40663</td>
<td>0.99993</td>
<td>0.10823</td>
<td>0.89170</td>
</tr>
<tr>
<td>0.41663</td>
<td>0.99993</td>
<td>0.09254</td>
<td>0.90738</td>
</tr>
<tr>
<td>0.44329</td>
<td>0.99993</td>
<td>0.10822</td>
<td>0.89171</td>
</tr>
<tr>
<td>0.46995</td>
<td>0.99993</td>
<td>0.09250</td>
<td>0.90743</td>
</tr>
<tr>
<td>0.48329</td>
<td>0.99993</td>
<td>0.10802</td>
<td>0.89191</td>
</tr>
<tr>
<td>0.53328</td>
<td>0.99993</td>
<td>0.09245</td>
<td>0.90748</td>
</tr>
<tr>
<td>0.54328</td>
<td>0.99993</td>
<td>0.10775</td>
<td>0.89218</td>
</tr>
</tbody>
</table>
Figure A1.3. Histograms for two different parameters ($\Delta X_{MB}$ and $\Delta R_{sep}$) for the CuAl$_2$ (tetragonal, I4/mcm) and KHg$_2$ (orthorhombic, Imma) structure types; the two structure types have almost same number of data entries in the LPF; where, $\Delta R_{sep} = 0.53328$ becomes the classifier (a vertical solid line) for the separation of the two structure types; the procedure to find the splitting value using information gain is described in the text.

A1.3.2. Recursive partitioning of a multi-class problem

Figure A1.4 shows the partitioning procedure of multi-dimensional parameter space
and the representation by classification tree. The collection of points denotes the 840 AB₃ compounds. To demonstrate how a structure domain is refined by the partitioning steps, the compounds with AlB₂-type structure were highlighted with red color as the example.

**Figure A1.4.** The partitioning of parameter space and the corresponding classification tree. The partitioning procedure (step a and b) is shown at the projection of ΔRₓₛᵖ-VE plane. The blue lines in the plots (left) indicate the hyperplanes which divide the parameter space. Red dots indicate the compounds with AlB₂ structure type. Tree structures (right) represent the position of the hyperplanes as the splitting condition. Red arrows indicate the conditions for the separation of AlB₂ structure domain. The numbers under each node represent the number of corresponding compounds divided by the node condition.

Out of 840 compounds, there are 45 AlB₂, 16 CaC₂, 21 CaF₂, ..., 15 ZrSi₂ type structures. At the root node (i.e., before partitioning), the entropy of the data, Hᵣ₀,
$H_{Root} = - \sum p(x) \log_2 p(x)$

$= - \left[ p(AlB_2) \log_2 p(AlB_2) + p(CaC_2) \log_2 p(CaC_2) + p(CaF_2) \log_2 p(CaF_2) + \ldots + p(ZrSi_2) \log_2 p(ZrSi_2) \right]$  

$= - \left[ (45/840) \log_2 (45/840) + (16/840) \log_2 (16/840) + (21/840) \log_2 (21/840) + \ldots + (15/840) \log_2 (15/840) \right]$  

$= 4.3751 \text{ bits}$

At each step of the partitioning, one AP value which minimizes the entropy of the data set, i.e. for the maximum IG, is chosen as the splitting parameter by exhaustively calculating the IG for all the possible AP values. At the first partition step, for instance, maximum IG = 0.4468 bits for VE, 0.3591 bits for $\Delta X_{MB}$, 0.5867 bits for $\Delta R_{zs+p}$, 0.1821 bits for $n_{av}$, 0.2805 bits for $\Delta X_{Pauling}$, 0.2975 bits for $\Delta \Phi^*$, and 0.2440 bits for $\Delta n_{ws^{1/3}}$ are achieved. Thus, $\Delta R_{zs+p}$ is selected as the splitting parameter. In practice, the entropy of two sub-nodes ($H_a$) is minimized at $\Delta R_{zs+p} = -0.03333$ and it becomes the first partitioning condition.

$H_a = \{P_1(H_{left \ node}) + P_2(H_{right \ node})\}$

$= \{(221/840)(3.4954) + (619/840)(3.893)\}$

$= 3.7884 \text{ bits}$

The information gain, $IG = H_R - H_a = 4.3751 - 3.7884 = 0.5867 \text{ bits}$

In the same way, at the second partition, IG is maximized at $VE=8.0002$

$H_b = \{(123/619)(2.6184) + (496/619)(3.5911)\}$

$= 3.3978 \text{ bits}$
IG = H_{a,\text{right}} - H_{b} = (619/840)(3.893 - 3.3978) = 0.3649 \text{ bits}

In this way, the classification tree is grown and then is pruned by the stopping procedure as explained. The following Figure A1.5 shows the whole procedure of the classification of crystal structure data.

**Figure A1.5.** The schematic procedure of the data classification by using recursive partitioning for the application to the crystal chemistry research.
References


APPENDIX II
CRYSTALLOGRAPHIC GROUP-SUBGROUP RELATIONS

A2.1. Space group symbols

Hermann-Mauguin symbol is the notation of the symmetry elements in the point
groups, plane groups, and space groups. The symbol consists of four character parts
as follows:

\[
\begin{align*}
1 & \quad 2 & \quad 3 & \quad 4 \\
\end{align*}
\]

The first character indicates the crystal lattice centering, that is,

- \( P \) = primitive
- \( A, B, \) and \( C \) = base-centered
- \( F \) = face-centered
- \( I \) = body-centered
- \( R \) = rhombohedral

Seven crystal systems with the lattice centering of the five different types consist of
14 Bravais lattices (Table A2.1.)

From the second to fourth parts denote the symmetry elements along certain
directions. Symmetry operations are categorized into the basic and combined
symmetry operations:

- Basic symmetry operations
  - Rotation axis : 1, 2, 3, 4, and 6-fold
  - Reflection : m, /m
- Inversion : \( i \)
- Translation : \( t \) (translation vector)

- Combined symmetry operations
  - Roto-inversion (= rotation + inversion) : \( \bar{1}, \bar{2}, \bar{3}, \bar{4}, \) and \( \bar{6} \)
  - Roto-reflection (= rotation + reflection) : \( 1m, 2m, 1/m, 2/m, \) etc.
  - Screw axis (= rotation + translation) : \( 2i, 3i, 4i, 6i, \) etc.
  - Glide plane (= reflection + translation) : \( a, b, c, n, d \)

Table A2.1. Crystallographic space groups.

<table>
<thead>
<tr>
<th>Crystal System</th>
<th>Crystal Lattice</th>
<th>Symmetry Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>C</td>
</tr>
<tr>
<td><strong>Triclinic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( a=b=c; \alpha+\beta+\gamma )</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Monoclinic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( a=b=c; \alpha=\gamma=90^\circ, \beta=90^\circ )</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Orthorhombic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( a=b=c; \alpha=\beta=\gamma=90^\circ )</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Tetragonal</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( a=b=c; \alpha=\beta=90^\circ )</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hexagonal / Trigonal</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( a=b=c; \alpha=\beta=90^\circ, \gamma=120^\circ )</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cubic</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For example, a hexagonal space group symbol, \( P6/m2/m2/m \), denotes a set of symmetry operations which includes the following symmetry information.
A2.2. Symmetry reduction between space groups

Space group is a set of symmetry operations. A space group $G_2$ becomes a subgroup of another space group $G_1$, if the set of symmetry operations of $G_2$ is the subset of those of $G_1$. In the same way, $G_3$ is a subgroup of $G_2$ when its symmetry-operations set is the subset for $G_2$.

$$G_1 = \{g_1, g_2, g_3, g_4, \ldots, g_{n-2}, g_{n-1}, g_n\}$$

$$G_2 = \{g_3, g_4, \ldots, g_{n-2}\}$$

$$G_3 = \{g_3, g_4\}$$

In such a case, it is said that the symmetry is reduced, and the symmetry relation of $G_1$ and $G_2$ is described using an arrow pointing from higher symmetry space group to the lower one. Here, $G_1$ becomes a supergroup of $G_2$, and $G_2$ becomes a subgroup of $G_1$. If there is no intermediate space group between $G_1$ and $G_2$, $G_2$ is called as a maximal subgroup of $G_1$. 
For three space groups with group-subgroup relations, the symmetry reduction can be described as follows.

There are two types of subgroups (=topologically equivalent) of a space group, which are called as t-type and k-type subgroups.

(i)  \textit{t}-subgroups: “lattice-equivalent” subgroups with the same translations. In this case, a subgroup $G_2$ contains all the translations of the super group $G_1$.

(ii) \textit{k}-subgroups: “class-equivalent” subgroups of the same class. In this case, $G_1$ and $G_2$ have the same crystal class (=point group) but belong to different space-group types. $G_2$ has lost translational symmetry; that is, the primitive cell corresponding to $G_2$ is larger than that of $G_1$.

One more thing should be mentioned is that the transformations of the coordinate system. When the unit cells of two crystal structures are transformed, the relations of the coordinate systems are defined by a transformation matrix which consists of the rotation and origin shift of the coordinates. The basis vectors of a new coordinate, $a'$, $b'$, $c'$ have the relations with those of the current coordinate, $a$, $b$, $c$ by a
transformation matrix $P$; that is,

$$(a', b', c') = (a, b, c)P$$

$$= (a, b, c) \begin{pmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{pmatrix}$$

For instance, suppose $c \rightarrow 1/2c$, the transformation matrix, $P$ is then

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/2 \end{pmatrix}$$

The origin shift is denoted by a triplet of numbers, e.g. 0, 1/4, 1/4. This means that the origin is shifted by 0, 1/4$b$, 1/4$c$ from the original coordinates.

References
