

INTERIM REPORT (SPECIAL PROJECT SPC 98-4028)

Contract No. F61775-98-WE013

covering period 15 March 1998 - 14 March 1999

**DEVELOPMENT OF A NEW TECHNIQUE FOR DISCOVERING
SYSTEMATICALLY HIDDEN PATTERNS**

By Pierre Villars

Material Phases Data System (MPDS), CH-6354 Vitznau, Switzerland

INTRODUCTION

Materials design is still mainly based on the in materials science known concepts and intuition of the experimentalists. Analyzing the conditions that make it possible to search for the in materials science known concepts shows that it was not a new technique, a unique experimental observation, or an abstruse theory which formed the take-off point. It was rather the amassing of a critical volume of experimentally determined data in the literature that permitted an individual with deep insight to perceive an underlying pattern not previously apparent.

Extending these facts to a new area of materials design leads to the following four key-points:

- I) The creation and the use of huge, critically evaluated materials databases which comprehensively covers the published world literature (materials databases).
- II) Computer-aided reduction of the elemental property parameters and systematic combinations of them to find the relevant 3D-feature sets which qualitatively can link materials properties with the chemical species present (semi-empirical approaches).
- III) Refinement and optimization of the qualitatively obtained results under II) with the help of neuro-computing leading to quantitative results (neuro-computing).
- IV) Focusing on predicted, most promising materials systems with the aim to reduce the experimental work for its verification, as well as trying to create a theoretical based explanation for such quantitative results (first principle calculations).

Materials Databases:

The amount of critically evaluated materials data starts to reach an acceptable volume, but is still far away from calling it comprehensive, below are the 6 most significant materials databases (available in electronic form) listed:

ICSD

This Inorganic Structure Database is maintained by the Fachinformationszentrum in Karlsruhe, Germany and contains crystallographic data for inorganic compounds.

CRYSTMET

This Intermetallic Structure Database was maintained until 1 April 1997 by National Research Council of Canada NRCC, CISTI, its hard copy versions are Pearson's Handbook of Crystallographic Data for Intermetallic Phases, by P. Villars and L. Calvert, ASM International, 1991 and Pearson's Desk, by P. Villars

Edition, ASM International, 1997 which contain crystallographic data for intermetallics and alloys.

ICDD PDF2

DTIC QUALITY INSPECTED 4

AQF99-10-1862

19990728 025

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE 1999	3. REPORT TYPE AND DATES COVERED Final Report	
4. TITLE AND SUBTITLE Development of a New Technique for Discovering Systematically Hidden Patterns			5. FUNDING NUMBERS F61775-98-WE013	
6. AUTHOR(S) Dr. Pierre Villars				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) MPDS 400 Schwanen Vitznau CH-6354 Switzerland			8. PERFORMING ORGANIZATION REPORT NUMBER N/A	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) EOARD PSC 802 BOX 14 FPO 09499-0200			10. SPONSORING/MONITORING AGENCY REPORT NUMBER SPC 98-4028	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.			12b. DISTRIBUTION CODE A	
13. ABSTRACT (Maximum 200 words) This report results from a contract tasking MPDS as follows: The contractor will investigate development of a fully-automated computerized database to predict critical properties of new materials. The research will use a 3-D (plus 4-D (color) and 5-D (shape)) parameterized "discovery space" to correlate atomic property expression of constituent chemical elements and compound/system properties using crystal structure data of 30,000 distinct compounds.				
14. SUBJECT TERMS EOARD, Materials, Chemistry			15. NUMBER OF PAGES 15	
			16. PRICE CODE N/A	
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UL	

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

This Powder Diffraction Patterns Database is maintained by the International Centre for Diffraction Data in Swarthmore, Philadelphia which contains mainly measured powder patterns.

BINARY ALLOY PHASE DIAGRAMS CD-ROM

This CD-ROM is maintained by ASM International (editor-in-chief: T.B. Massalski)

TERNARY ALLOY PHASE DIAGRAMS CD-ROM

This CD-ROM is maintained by ASM International (editor-in-chief: P. Villars)

LPF (LINUS PAULING FILE)

A basic database for alloys, intermetallics and inorganics. This file is now in the process to be build up by the Japan Science and Technology Corporation JST in Tokyo and MPDS in Switzerland and is planned to enter the yearly update stage in 2007 which covers structure, diffraction, property and constitution data (editor-in-chief: P. Villars)

Semi-empirical Approaches:

There exists in the world literature a whole range of 'highest quality' correlations between materials properties and the chemical species present. To all of them is common that they were found by semi-empirical approaches based on a *small to large amount of data*. A comprehensive review is given in the book "Intermetallic Compounds, Principles and Practice (Volume 1), chapter 11 'Factors Governing Crystal Structures' by P. Villars; edited by J.H. Westbrook and R.L. Fleischer. published by John Wiley & Sons (1995) ISBN 0-471-94219 7. Here we propose the development of a fully automated discovery space to search for the relevant 3D-feature sets (derived from elemental property parameters of the chemical species present) to correlate qualitatively materials properties with the chemical species present.

Neuro-computing:

After such qualitative correlations have been discovered the following neuro-computing approaches showed to be in several examples surprisingly striking in improving the qualitative correlations clearly towards quantitative correlations. The following approaches showed to be very successful:

- Function Approximation - Ensemble Approach (Dr. Igel'nik)
- Function Approximation - Orthogonal Approach (Dr. Chen)
- Function Approximation - Auto-Associative Filtering (Dr. Thaler)
- Clustering and Visualization (Dr. Pao)

First principle calculations:

Even after establishing above mentioned quantitative correlations the number of to be experimentally verified predictions is in general too high because most advanced materials of today's interest are ternary or quaternary systems. In the case of ternary systems there exists 161'700 and for quaternary systems 3'921'225 potentials. In addition, in order to establish structures and phase relationships, one has to prepare and investigate at least 10 times more samples per system by going e.g. from ternary to quaternary. It is therefore a need to focus first principle calculations towards selected groups of most promising compound classes before starting with experimental verification. In the best case one can achieve to create a theoretical based explanation for such quantitative results from first principle calculations, which would be essential to find the direct processing parameters to optimize its production.

WORK REPORT

Materials Databases:

In this report we show on the example of ternary formers respectively nonformers how successful the collaboration/interplay between materials databases, semi-empirical approaches and neuro-computing can be. Here we used the data of the following databases:

PEARSON'S DESK EDITION by P. Villars (electronic version), ASM International, 1997 which contain crystallographic data for intermetallics and alloys.

BINARY ALLOY PHASE DIAGRAMS CD-ROM

This CD-ROM is maintained by ASM International (editor-in-chief: T.B. Massalski)

TERNARY ALLOY PHASE DIAGRAMS CD-ROM

This CD-ROM is maintained by ASM International (editor-in-chief: P. Villars).

and we found nonformers/formers infos as follows:

NONFORMERS

a) 1231 (from *TERNARY ALLOY PHASE DIAGRAMS CD-ROM*, after inspection by PV: 508 of them not decided, 130 switched from nonformers to formers)

= 596 nonformers

b) 2327 (from *BINARY ALLOY PHASE DIAGRAMS CD-ROM* derived from 676 binary nonformers)

= 2327 nonformers

c) 244 (from *TERNARY ALLOY PHASE DIAGRAMS CD-ROM UPDATES*, after inspection by PV: 133 not decided)

= 111 nonformers

total: 3034 nonformers

FORMERS

d) 130 (switched from the 1231 nonformers to formers, see a))

= 130 formers

e) 4104 (from *PEARSON'S DESK EDITION* of systems having at least 1 compound with a 'ternary structure type')

= 4104 formers

f) 705 from Pearson's HB crystallizing in the structure type $\text{Al}_4\text{Ba}(\text{CrTh}_2\text{Si}_2)\text{tI10,139}$

= 705 formers

g) 343 (from *TERNARY ALLOY PHASE DIAGRAMS CD-ROM UPDATES*, after inspection by PV: 242 not decided)

= 101 formers

total: 5040 formers

total: 3034 nonformers + 5040 formers = 8074 infos on distinct ternary systems

After eliminating overlaps we end up with 4 datasets (created by Al Jackson):

set 1: **1067 nonformers** used as <<data1test.txt>>

set 2: **2327 nonformers** used as <<data2train.txt>>

set 3: **4031 formers** used as <<data3train.txt>>

set 4: **244 nonformers/formers** used as <<244dataset4_rev.txt>>

total: 7669 ternary systems (405 overlaps excluded)

Semi-empirical Approaches:

As the starting point for the search of the relevant 3D-feature sets are always the elemental property parameters it is important to know which elemental property parameters are least dependent from each others and having access to complete data-sets for all known elemental property parameters. Therefore the first work was to collect and evaluate as many as possible of such elemental property parameters data-sets and group them into distinct independent different groups, here called factors. Elemental property parameters belonging to the same factor are most similar to each others. As starting point I reviewed the literature and found over 300 elemental property parameters data-sets, most of them having data for less than 80 % of all chemical elements (the ones with less than 80 % I excluded as they would be of very limited practical use). Below I have grouped the 43 elemental property parameters data-sets (for which each set has at least for 92% of the 100 chemical elements published values) into 6(7) distinct factors, which are most dis-similar from each others. Within the 6(7) factors, if there are more than one elemental property parameters data-set they are very similar to each others.

We made a very interesting observation:

**that the 6(7) factors have units corresponding to
the 7 SI fundamental units (m, K, mol, A, s, kg, cd),**

these cannot be derived from each others, all the other SI units are derived from one of those.

Based on that observation I am convinced that the elemental property parameters data-sets with more complex units are actually derived from those, and therefore dependent from each others, most obvious is the density ρ [kg m⁻³]. All these I have group under derived factors. I believe I found a new additional factor to the already previously by us known 5 factors, the FREQUENCY FACTOR, and we predict the existence of an additional factor, the OPTICAL FACTOR.

Here are the 6(7) factors listed:

- **SIZE FACTOR** = (SI fundamental unit for SPACE: m)
- **HEAT FACTOR** = (SI fundamental unit for HEAT: K)
- **ELECTRO-CHEMICAL FACTOR** = (SI fundamental unit for MOLECULAR PHYSICS: mol)
- **VALENCE ELECTRON FACTOR** = (SI fundamental unit for ELECTRICITY/MAGNETISM: A)
- **ATOMIC NUMBER FACTOR** = (SI fundamental unit for MECHANIC: kg)
- **FREQUENCY FACTOR** = (SI fundamental unit for TIME: s)
- **OPTICAL FACTOR** = (SI fundamental unit for OPTICAL RADIATION: cd)

Below are the 43 elemental property parameters data-sets grouped according to 6(7) factors:

SIZE FACTOR = (SI fundamental unit for SPACE: m)

SF1 RADI PSEUDO-POTENTIAL (Zunger) [a.u.]

SF2 radii ionic (Yagoda) [A]

SF3 radii covalent [pm]

SF4 radii metal (Waber) [A]

SF5 distance valence electron (Schubert) [A]

SF6 distance core electron (Schubert) [A]

SF7 volume atom (Villars, Daams) [nm³]

SF8 atomic environments (Villars, Daams) [different polyhedra]

Villars, 1995

Samsanov, 1966

Chen, 1997

Pearson

Schubert

Schubert

Villars, Daams, 1995

Villars, Daams, 1995

HEAT (formerly called COHESION-ENERGY) FACTOR = (SI fundamental unit for HEAT: K)

HF1	TEMPERATURE MELTING [K]	Villars, 1995
HF2	temperature boiling [K]	Emsley, 1991
HF3	enthalpy vaporization [kJ mol ⁻¹]	Emsley, 1991
HF4	modulus compression [m ² N ⁻¹]	Kittel, 1980
HF5	energy cohesive (Brewer) [eV=kJ mol ⁻¹]	Kittel, 1978
HF6	enthalpy melting [kJ mol ⁻¹]	Chen, 1997
HF7	enthalpy atomization [kJ mol ⁻¹]	Emsley, 1991

ELECTRO-CHEMICAL FACTOR = (SI fundamental unit for MOLECULAR PHYSICS: mol)

ECF1	ELECTRONEGATIVITY (MARTYNOV&BATSANOV) [/]	Villars, 1995
ECF2	electronegativity (Pauling) [/]	Chen, 1997
ECF3	electronegativity (Alfred-Rochow) [/]	Chen, 1997
ECF4	electronegativity absolute [/]	Chen, 1997
ECF5	energy ionization first [J mol ⁻¹]	Chen, 1997
ECF6	energy ionization second [J mol ⁻¹]	Chen, 1997
ECF7	energy ionization third [J mol ⁻¹]	Chen, 1997

VALENCE ELECTRON FACTOR = (SI fundamental unit for ELECTRICITY/MAGNETISM: A)

VEF1	VALENCE ELECTRON NUMBER [A]	Villars, 1995
VEF2	group number [/]	periodic table

ATOMIC WEIGHT (formerly called ATOMIC NUMBER) FACTOR = (SI fundamental unit for MECHANIC: kg)

AWF1	WEIGHT ATOMIC [kg]	Emsley, 1991
AWF2	atomic number [/]	periodic table
AWF3	quantum number [/]	periodic table
AWF4	charge nuclear effective (Clementi) [/]	Emsley, 1991
AWF5	atomic electron scattering factor [/] for ...=0.5	Samsanov
AWF6	normal atomic configuration (Thaler)	Busch, 1988

FREQUENCY FACTOR = (SI fundamental unit for TIME: s)

FF1	MAGNETIC RESONANCE [Hz= s ⁻¹]	Kaye, Laby, 1990
FF2	magnetic frequency of nuclei [Hz]	Emsley, 1991

OPTICAL FACTOR = (SI fundamental unit for OPTICAL RADIATION: cd) ?**DERIVED FACTOR**

DF1	density [kg m ⁻³]	Emsley, 1991
DF2	mass attenuation coefficient for .. radiation m ² kg ⁻¹]	International tables
DF3	electrochemical weight equivalent [kg K ⁻¹]	Samsanov
DF4	entropy of solid [J mol ⁻¹ K ⁻¹]	Samsanov
DF5	molar heat capacity [J mol ⁻¹ K ⁻¹]	Chen, 1997
DF6	Mendeleev number (Pettifor) [/]	Pettifor
DF7	thermal neutron capture cross section [barns]	Emsley, 1991
DF8	moment nuclear magnetic [J T ⁻¹]	Kittel
DF9	spin nuclei [J s]	Kittel
DF10	oxidation state first [/]	Chen, 1997
DF11	electron affinity [kJ mol ⁻¹]	Emsley, 1991

(the first 6 having very regular, symmetrical appearance in a elemental property parameters vs. periodic table plot, the last 5 look very 'wild' (but their values seems too have high accuracy))

Because of the importance of the elemental property parameters data-sets for the discovery of the relevant 3D-feature sets I propose to investigate the following dependences:

1) I propose that we take for each of the 6(7) factors my recommended elemental property parameters data-set and check through neuro-computing that they are really most independent. The recommended elemental property parameters data-sets are:

- SIZE FACTOR: *RADII PSEUDO-POTENTIAL (Zunger) [a.u.]*
- HEAT FACTOR: *TEMPERATURE MELTING [K]*
- ELECTRO-CHEMICAL FACTOR: *ELECTRONEGATIVITY (M&B) [/]*
- VALENCE ELECTRON FACTOR: *VALENCE ELECTRON NUMBER [A]*
- ATOMIC NUMBER FACTOR: *WEIGHT ATOMIC [kg]*
- FREQUENCY FACTOR: *MAGNETIC RESONANCE [Hz=s-1]*
- OPTICAL FACTOR: ? [cd]

2) Check through neuro-computing the similarity of the elemental property parameters data-sets within each factors and if one stands out as the 'best'.

3) Check through neuro-computing if you can correlate the 11 elemental property parameters data-sets under derived factor with combinations of the 6(7) factors, and if you can come up with my 'predicted' elemental property parameters data-set for the optical factor?

4) Do you have an idea how we can include:

- * atomic environments (Villars, Daams) [different polyhedra] Villars, Damms, 1995
- * normal atomic configuration (Thaler) Busch, 1988

which represent the only two elemental property parameter data-sets with non-numerical values?

After we have established the 6(7) factors and have chosen the best elemental property parameters data-set for each factor we have to generate systematically all combinations, which would represent all potential 3D-features sets. For that purpose we have developed a code called DISCOVERY.

In the until now performed neuro-computing work we have only used the following elemental property parameters data-sets:

- *RADII PSEUDO-POTENTIAL (Zunger) [a.u.]*
- *TEMPERATURE MELTING [K]*
- *ELECTRONEGATIVITY (Martynov and Batsanov) [/]*
- *VALENCE ELECTRON NUMBER [A]*
- *ATOMIC NUMBER [Z]*

and the following feature sets:

- *VILLARS' FEATURE SET* (Taking the average sum instead the difference in Villars's features gives better results, Chen):
- element_A
- element_B
- element_C
- RADII PSEUDO-POTENTIAL (Zunger) difference (average sum)
- TEMPERATURE MELTING ratio
- VALENCE ELECTRON NUMBER difference (average sum)
- ELECTRONEGATIVITY (Martynov and Batsanov) difference (average sum)
- ATOMIC NUMBER ratio
- *THALER'S FEATURE SET* (normal atomic configuration):
- core_A, core_B, core_C
- sA, sB, sC
- pA, pB, pC
- dA, dB, dC
- fA, fB, fC

Some explanations to the software DISCOVERY:

We divided the DISCOVERY code into three tasks as follows:

- 1) Building an automatic generator for 3D-feature sets resulting from combinations of *elemental property parameters* and *mathematical operations*.
- 2) Building a program for detection of those 3D-feature sets with the "best" separation of the material property, which is investigated.
- 3) Building a sophisticated 3D-Graphics program for the interactive investigation of the best 3-D feature sets.

So far item 1) was a relatively simple task, item 2) meanwhile comes to reasonable results and item 3) is under development. We decided to build two separate modules, one containing a 3D-graphics interface, and one containing a number crunching kernel running in background (we call it the „QFD Kernel“).

1) The number of 'elemental property parameters' and 'matemetical operations' have to be chosen very carefully because otherwise the total number of combinations becomes really „astronomic". Six elemental property parameters and five operations gives 30 combinations. Three of them each gives one 3D-feature set to be investigated. That means $(30 \times 29 \times 28) / (3 \times 2) = 4060$ 3D-feature sets. One elemental property parameter more would give 6545 3D-feature sets, one operation more would give 7140 3D-feature sets, and one elemental property parameter and one operation more would give 11480 3D-feature sets.

2) The central problem is: how to decide by a program, what is a „good" separation and what is not. What humans normally do using their eyes and brains is: finding some kind of border between areas and then counting how many points with the same property are at the same side of this border. But finding such a border is a very complex task for a program. In one dimensions finding a border point which seperates „left" and „right" points is relatively easy. In two dimensions you have to find a more or less complex border line between the areas and this is already a complex optimization problem. But in three dimension you would have to find a more or less complex surface, and though algorithms to find them exist, they would take too much calculation time to be applied on thousands of 3D-feature sets. So the detection algorithm must work on a much simpler basis. But we must be aware, that this strategy leads to different definitions of what is the quality of separation within one 3D-feature set. In other words: „Separation of 97,92%" - What does that mean? It has to be pointed out before comparing results of different QFD strategies.

However, our starting point is: If a separation is „good" then many points' next neighbour(s) must have the same system property. And this can be detected with simple distance calculations - though it took a lot of work to minimize the total number of calculations. Imagine, that in principle all distances from each point has to be calculated to find out, what is the shortest and therefore what is the next neighbour. And 6358 points gives $6358 \times 6357 / 2 = 20.7$ million distances - for each 3-D feature set! We were able to reduce that to approx. 100,000 calculations for a rough analysis and approx. 1 million calulations for a more precise analysis. But nevertheless the analysis is in principle „afflicted" with uncertainties - some neighbours could not be found and therefore the analysis result could be more bad than it has to be. These uncertainties are about 0.1% (for the more precise analysis) and I think they can be ignored (but has to be mentioned).

3) Since the graphics part of QFD is still under development, we had to visualize only some results using the current DIAMOND version. Below are some results for the 6358 nonformers/
formers: The calculation and analysis of 6040 3D-feature sets takes about 30 minutes on a Pentium

133 / 32 Mbytes RAM and about 10 minutes on a Pentium II 300 / 64 Mbytes RAM. I think this is a reasonable time even to investigate some more 3-D feature sets. As one can see from the QFD Kernel Report the best result gives 6297 Hits = 97.92% Separation and there are a lot of very „good" results. But note, that this only means: 97.92% of all points have a next neighbour with the same system property. So what about the „other" neighbours? A really good separation means that points of the same colors give clusters, which are as big as possible. Therefore the QFD Kernel can investigate the best results again and more detailed. It's no surprise that the number of „Hits" decreases with increasing number of considered next neighbours because even large clusters will come to an end sooner or later. This indicates the real quality of separation, because a good separation starts with a high number of Hits and decreases as slow as possible. That means, that not always the run with the highest starting point ("best" result) gives the very best separation. Fig. 1 shows the run with the highest number of first next neighbours with the same system property (97,9%). One can see, that in fact the separation is very good. Fig. 2 shows a run with a rather high number of first next neighbours with the same system property (96.6%) but considering more next neighbours the number decreases rapidly. Although almost each point is surrounded by points with the same color, the clusters in total are small and mixed randomly, and the separation is poor.

This can be best demonstrated on a practical example.

Considering the following 6 elemental property parameters:

- 1) atomic number
- 2) electronegativity after M&B
- 3) magnetic resonance
- 4) temperature melting
- 5) radii pseudo-potential after Zunger
- 6) valence electron number

Considering the following 5 mathematical operations:

- 11) sum
- 12) difference
- 13) ratio
- 14) product
- 15) maximum

Considering 6358 ternary systems

4031 formers +

(from *PEARSON'S DESK EDITION*)

2327 nonformers

(derived from 676 binary nonformers, from *BINARY ALLOY PHASE DIAGRAMS CD-ROM*)

Below are listed the results for the 20 'best' from all

$(30 \times 29 \times 28) / (3 \times 2) = 4060$ 3D-feature sets (see figs 1 + 2):

This means e.g. for 6358 ternary formers/nonformers infos $6358 \times 6357 / 2 = 20'208'903$ distances for each 3D-feature set.

1	01-15 / 05-12 / 05-15	6297 (97.92%)
2	02-15 / 05-12 / 05-15	6288 (97.78%)
3	02-12 / 05-12 / 05-15	6282 (97.68%)
4	01-12 / 05-12 / 05-15	6276 (97.59%)
5	01-12 / 01-15 / 05-14	6268 (97.47%)

6	01-15 / 05-12 / 05-13	6261 (97.36%)
7	02-15 / 05-12 / 05-13	6257 (97.29%)
8	01-11 / 05-12 / 05-15	6256 (97.28%)
9	02-12 / 05-13 / 05-15	6250 (97.19%)
10	01-12 / 05-13 / 05-15	6249 (97.17%)
11	01-12 / 02-15 / 05-14	6248 (97.15%)
12	05-12 / 05-15 / 06-15	6245 (97.11%)
13	05-12 / 05-13 / 05-15	6245 (97.11%)
14	02-15 / 05-12 / 05-14	6242 (97.06%)
15	02-11 / 05-12 / 05-15	6241 (97.05%)
16	02-14 / 05-12 / 05-15	6241 (97.05%)
17	01-15 / 05-12 / 05-14	6239 (97.01%)
18	01-15 / 05-13 / 05-14	6236 (96.97%)
19	02-12 / 05-12 / 05-13	6236 (96.97%)
20	01-15 / 05-13 / 05-15	6235 (96.95%)

neuro-computing:

Using as starting point:

- a) the above summarized data on ternary formers/non-formers (info on about 8'000 chemical systems)
- b) the 5 elemental property parameters data-sets
- c) the feature sets according to Villars and according to Thaler
- d) the following four neuro-computing approaches were used
 - Function Approximation - Ensemble Approach (Dr. Igel'nik)
 - Function Approximation - Orthogonal Approach (Dr. Chen)
 - Function Approximation - Auto-Associative Filtering (Dr. Thaler)
 - Clustering and Visualization (Dr. Pao)

The under d) listed different neuro-computing approaches were used to find a quantitative correlation between the data ternary formers non-formers (see a)) and Villars+Thalers feature sets.

The most outstanding results were achieved using the following data sets for training:

set 2: **2327 nonformers** used as <<data2train.txt>>

set 3: **4031 formers** used as <<data3train.txt>>

Using data for 1631 systems training system with 4031 (formerly 4104, but 73 were overlapping) formers and 2327 non-formers (from binaries) gave 1608 systems correctly classified and 23 mis-classified systems. Below I summarize the results by checking the 23 mis-classified systems against the original literature trying to find errors, inconsistencies, etc.

The prediction results follow: (formers = 1, nonformers = 0)

System	cluster prediction	literature estimate	(re-evaluated literature estimate)	Comments of Re-Evaluation
Ag_Sr_Zn	0.282->0	1 (1)	(Ag9Sr10Zn),Ag9Sr10Zn,oP24,31 [R=0.069]	no pd
Al_B_Cu	-0.439->0	1 (?)	(AlB25Cu)AlB25Cu,tP58,118 [R=0.05] (AlB52Cu)AlB52Cu,hR113,166[R=0.06]	no pd
As_Hg_Pd	0.494->0	1 (?)	(AsHgPd5)AsPd5Ti,tP7,123 [two-phase sample]	no pd
B_In_Ni	-0.035->0	1 (1)	(BIn2Ni7)CaO3Ti,cP5,221 (B4InNi15)C6Cr23,cF116,225	1073K 2 ternaries
C_Mo_Re	0.171->0	1 (0)	(CMo3Re2)Al2CMo3,cP24,213-- >Mn,cP20,213	673K+1773K 0 ternary
Co_Cr_W	0.253->0	1 (0)	(Co23Cr15W15)Co5Cr2Mo3,hR53,148	300K+873K+973K 0 ternary
Cu_N_Pd	0.358->0	1 (?)	(Cu3NPdx,x=0.2-1)Cu3NPd,cP5,221	no pd
Fe_Ti_Y	0.198->0	1 (?)	(Fe16Ti3Y2)CeMn6Ni5,tP24,127 [R]	no pd
Ga_Sn_V	0.162->0	1 (1)	(GaSn2V2)GaSn2V2,oC40,64 [R=0.044]	no pd
Hf_Mo_Ni	0.134->0	1 (0)	(Hf9Mo4Ni)BHf9Mo4,hP28,194 [R=0.14, OXYGEN STABILIZED]	no pd
Li_Mg_Si	0.409->0	1 (1)	(Li8MgSi6)Li8MgSi6,mP48,11 [R=0.035]	no pd
Li_Na_Si	0.466->0	1 (1)	(Li3NaSi6)Li3NaSi6,oP40,62 [R=0.090]	no pd
Pt_Sb_Si	0.102->0	1 (?)	(Pt5SbSi)AsPd5Ti,tP7,123 [two-phase sample]	no pd
Sb_Sn_Zn	0.188->0	1 (0)	(Sb2SnZn)CuFeS2,tI16,122 + SZn,cF8,216	300K, Solidus 0 ternary
Si-Ta-Te	0.369->0	1 (1)	(SiTa4Te4)SiTa4Te4,oP36,55 [R=0.061]	no pd
Ag_Cr_U	1	0 (0)	no structural data	binaries N/N/N no pd
Bi_Ge_Sb	0.537->1	0 (0)	no structural data	binaries N/N/N no pd
Cr_Cu_V	0.816->1	0 (0)	no structural data	binaries N/N/N no pd
Cr_Mo_Pb	0.612->1	0 (0)	no structural data	binaries N/N/N no pd
Cu_Li_Na	0.569->1	0 (0)	no structural data	binaries N/N/N no pd
Fe_Li_Mn	1	0 (0)	no structural data	binaries N/N/N no pd
In-Ta-Tl	1	0 (?)	no structural data	binaries InTa3?/N/N no pd
V_W_Y	0.704->1	0 (0)	no structural data	binaries N/N/N no pd

13 of the 23 classified as '1' or '0' from literature estimates have been RE- CONFIRMED (and most likely the data are correct and experimentally established with highest quality data), the remaining 10 were either wrongly classified or are questionable (because of contradictory work).

- 1 (1) literature estimate re-confirmed, the existence of at least one ternary compound is established by highest quality data
- 0 (0) literature estimate re-confirmed, but non-existence of ternary is based on its established binary boundary diagrams
- ? (1) 1 literature estimate not re-confirmed, existence of ternary is questionable (contradiction/low quality data)
- ? (0) literature estimate not re-confirmed
- 1 (0) literature estimate was wrong

Therefore Pao's results are:

Total number of systems used for testing: 1631

Mis-classified systems: 13

systems classified correctly: 1618

Correct percentage: 99.2

instead of

Total number of systems used for testing: 1631

Mis-classified systems: 23

systems classified correctly: 1608

Correct percentage: 98.59

With that we reached over 99% prediction accuracy!!

SUMMARY CONCLUSIONS *from the data part of view*

- 1) The 4031 (formerly 4104, but 73 were overlapping) formers based on at least on ternary compound with a ternary crystal structure is of very high confidence level and therefore this selection criteria for formers is correct and simple.
- 2) The 2327 non-formers based on the existence of their three experimentally determined binary boundary phase diagrams (non-compund formers) is of very high confidence level and therefore the "Assumption that a ternary system is a nonformer when all its three binary boundary systems are also non-formers)".
- 3) The difference between a compound-forming (formers) and a non-compound-forming system (non-formers) is given by the fact that in the compound-forming system that at least one ternary compund has to be separated by three two-phase regions to its three adjacents binary and/or ternary compounds (in case where no phase phase diagram is known, A-B-C compounds with ternary crystal structures belong most often to the above mentioned case, in contrast to the pseudo-binary compounds).
- 4) Villars's or Thalers' features are adequate because accuracy of 95.8-99.2 can be achieved. Taking the average sum instead the difference in Villars's features gives better results. The results of DISCOVERY shows that even better results will be achieved with neuro-computing taking the 20 best 3-D feature sets as staring point.
- 5) The 13 of the 23 are less likely to be wrong, because the 6 formers of them (+) have its crystal structure have been refined by excellent researchers and the 7 non-formers (-) are based on our "Assumption that a ternary system is a nonformer when all its three binary boundary systems are also non-formers)", which gave an accuracy of > 99%.
- 6) The 1231 non-formers based on one isothermal section without ternary compound are inthead splitted into, and therefore their quality very 'limited':
 - 467 correctly classified non-formers
 - 127 wrongly classified non-formers (--> to be experimentally checked)
 - 130 formers (before wrongly grouped)
 - 507 formers/non-formers not to be decided by published experiments (to be checked by own experiments)

LONG-TERM-APPROACH

A comprehensive long-term approach to build up an international information- knowledge forum has to be able to solve the following sub-problems with the final aim to predict successfully materials with pre-defined intrinsic properties. The sub-problems are supported from observations made on large quantities of materials data:

1) To predict crystal structures of nonformers is in principle non-sense, therefore one has first to be able **to predict formers/nonformers with a high accuracy** [This is why we are working at the moment on that problem as a kind of test case].

2) As soon as one likes to make the step to compounds one has to introduce the concentration as variable in the feature sets space. In addition one has to define the boundary conditions in such a way that the step from binary to ternary to quaternary can be made smoothly using the same feature set, as well as treat binary, ternary and multinary together. This will minimize the number of sub-spaces and maximize the number of experimental facts.

3) Reduction the number of crystal structures, as there exists about 5-6'000 experimentally found crystal structures. By describing the periodic arrangements of the atoms from the light of the atomic environment (CN) of each atom in the crystal structure instead from the light of symmetry one can reduce the 5-6'000 to about 10%. This means there exists about 5-600 geometrically different crystal structures, here called atomic environment types (AET). The 'classical' structure types belonging to the same AET are structurally very similar and the gross feature of the structure is for all of them the same.

4) For most intrinsic material properties there exists a "NECESSARY BUT NOT SUFFICIENT" correlation to its crystal structure. Therefore the crystal structure and/or atomic environment type AET should be used as key to its intrinsic properties.

IN SHORT TO ACHIEVE AN ACCURATE 'VIRTUAL MATERIALS DESIGN' CAPABILITY

we need:

INTERACTIONS BETWEEN

- 1) access to huge materials databases
(*high-quality data*)
- 2) method to systematically discover the relevant correlation between elemental property parameters (3D-feature sets) and materials property
(*semi-empirical approach*)
- 3) Neuro-computing to optimize results
and make highest accuracy predictions >99%
(*neuro-computing*)
- 4) First principal calculations to create a theoretical based explanation for such quantitative results, to get a 'handle' on the processing control of its 'production of the wanted materials'
(*first principle calculation*)

FOLLOWING PREDICTION STEPS

*Being able to predict materials properties
one has to obey the following sequence:*

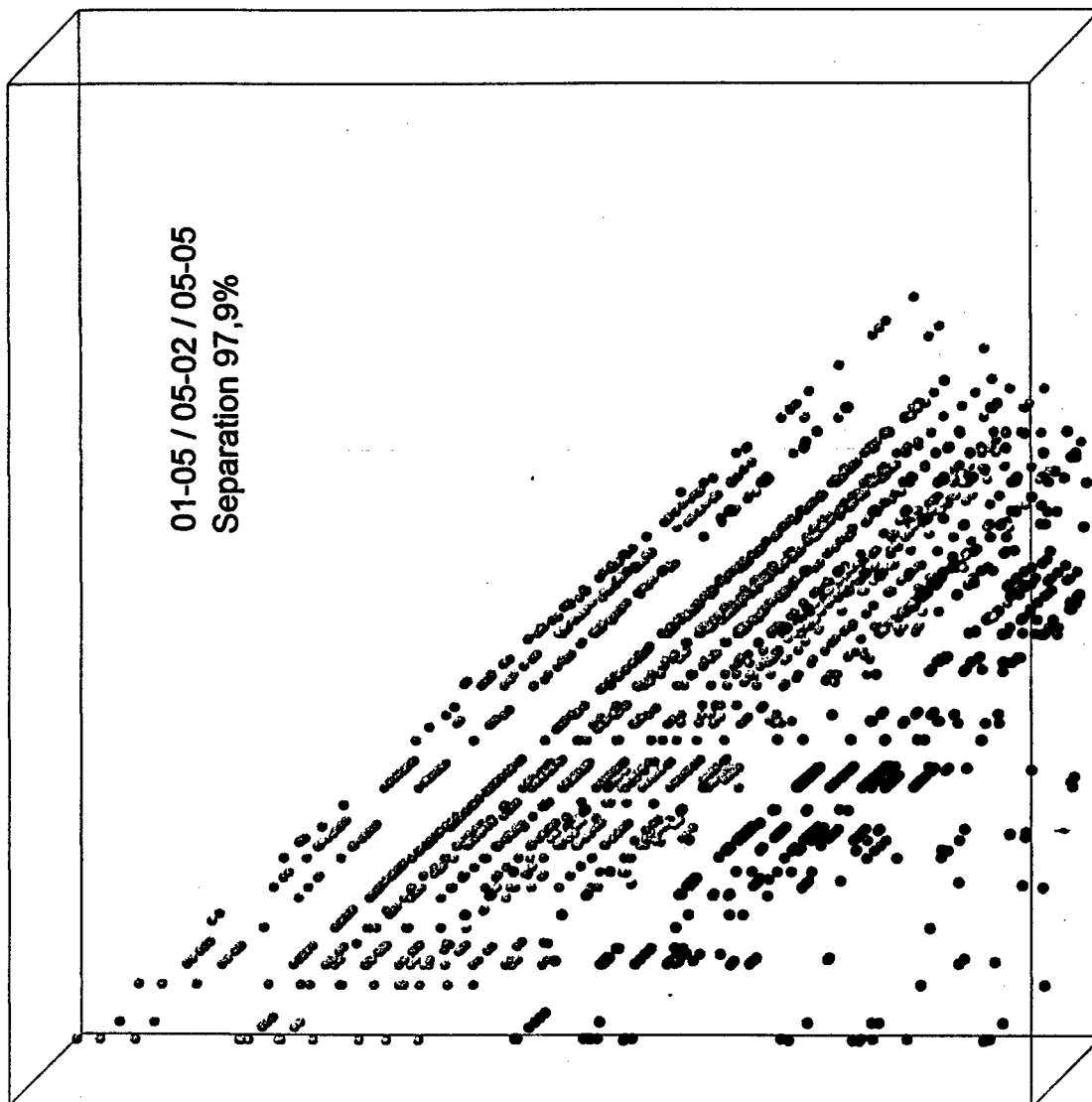
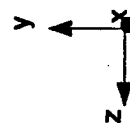
- 1) formers versus nonformers
- 2) stable compositions within the formers
- 3) crystal structure of the stable compositions within the formers
- 4) correlate materials properties with its crystal structure

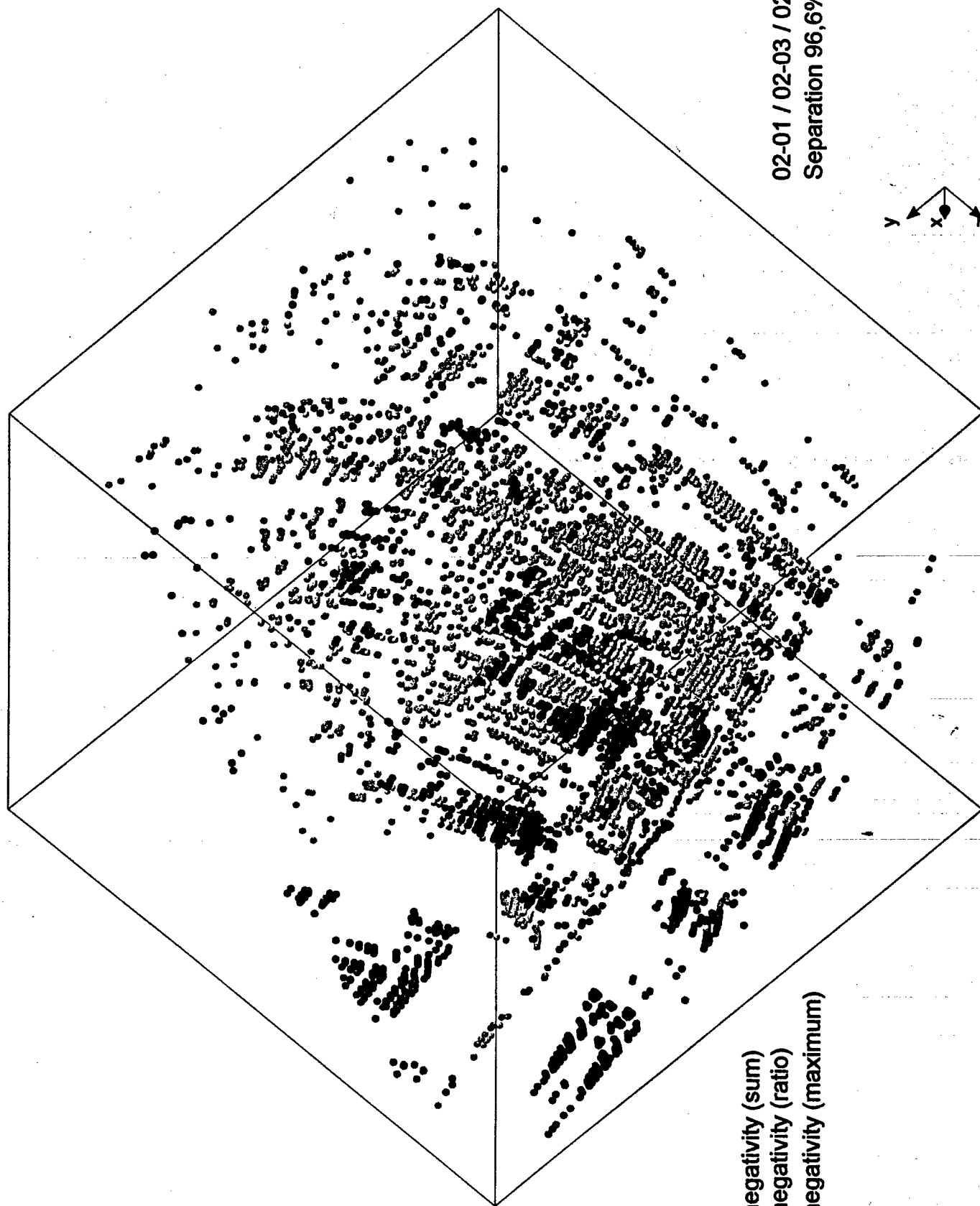
WATCH THE FOLLOWING

- 1) Do not miss any elemental property parameters (3D-feature sets)
- 2) Minimize the number of prediction steps by covering large materials groups
(e.g. treat binary, ternary, quaternary systems together instead just binaries at a fixed stoichiometry, this means increase data amount as much as possible!)
- 3) Maximize the accuracy of the prediction of each prediction step to over 99 %
(overall accuracy is most important)
- 4) Reduce the number of crystal structures by grouping structures having the same gross-feature, by moving from the 'classical' space group description to the atomic environment description
(at present about 6'000 structure types have been published)
- 5) Correlate materials properties with its crystal structures
(be aware that this is a necessary but not sufficient condition)
using the atomic environment description(e.g. most materials properties are found in less than 24
'classical' different crystal structures)

01-05 / 05-02 / 05-05
Separation 97,9%

x = atomic number (maximum)
y = pseudo-potential radius (differenc
z = pseudo-potential radius (maximum





02-01 / 02-03 / 02-05
Separation 96,6%

= electronegativity (sum)
= electronegativity (ratio)
= electronegativity (maximum)