Data Patterns and Links to Materials Theory: Theoretical Foundations for Heuristic Pattern Detection

Kim F. Ferris
Pacific Northwest National Laboratory

email: kim.ferris@pnl.gov • phone: (509) 375-3754 • fax: (509) 375-2290

Bobbie-Jo M. Webb-Robertson
Pacific Northwest National Laboratory

Dumont M. Jones
Proximate Technologies, LLC

Tutorial Outline

- Context relating to other presentations – before and after
- Goals
- Why pay attention?
- Bookkeeping – Contacts, Relevant Publications, Glossary
- Information architecture
  - Data
  - Models
  - Role of descriptors in model development
- Structure diagrams
  - Mathematical context of structure diagrams
    - Structure/loading plots
  - Historical perspective on structure diagrams
    - Pettifor plot original/reformatted data
- Levels of data
    - PCA/LDA diversity –Bulk Modulus example
    - First Principles Information

Reviewing Context from CALculation of PHase Diagrams (CALPHAD)

- Thermodynamic data and phase descriptions are fundamentally linked to the crystal structures – hence information linked to a crystallographic database is inherently useful.
- Data mining and statistical tools can be used to search for data and to pre-process the data in preparation for the CALPHAD-type optimizations.
- Crystal structures are an underlying basis of ab initio calculations, and address holes in the experimental data-sets.

Projecting Context for Data Mining

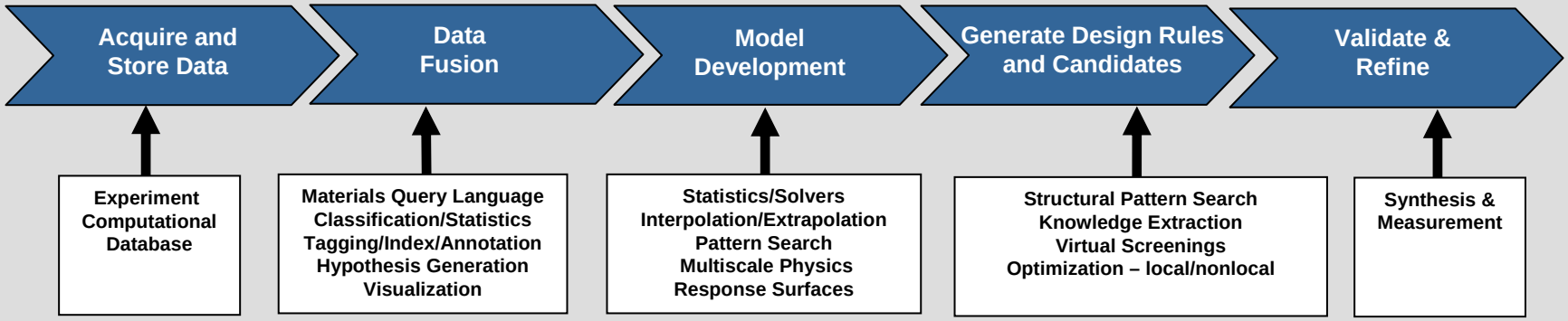
- Thermodynamic, energetic, crystal phase is fundamentally linked to materials behavior.
- No database is ever totally complete or current.
- Data mining/knowledge extraction tools can be used to search for property mappings and design rules.
- Model-driven exploration combined with high-throughput experiments accelerates discovery

Goals for this Section

Appreciation for:

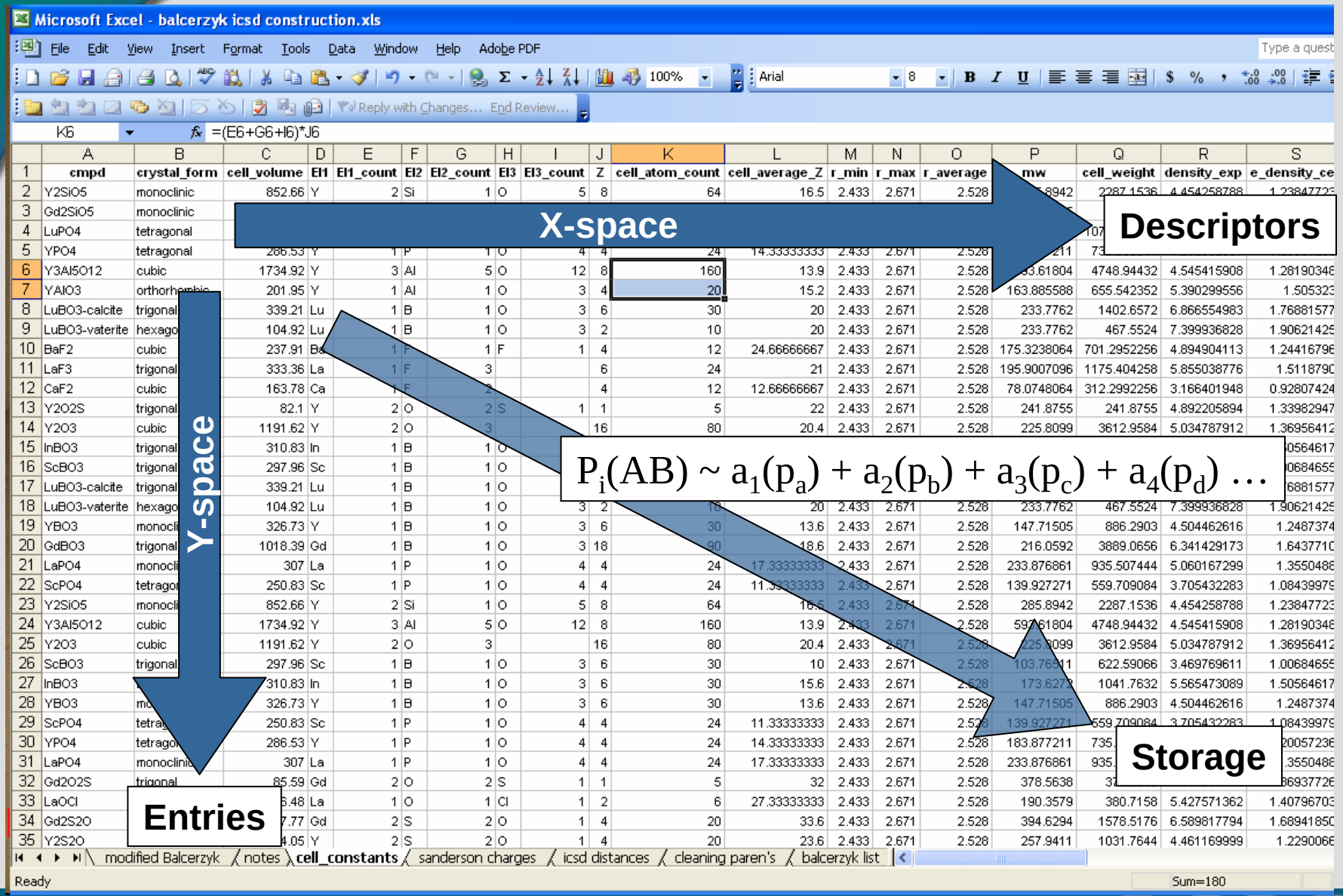
- Information architecture
  - Structure Maps (example: crystal structure)
  - Model Development
    - Need for formal knowledge extraction methods
    - Different Levels of Information
      - Data diversity
      - First principles electronic structure
      - Experimental measurement

Information Architecture



Want to combine information from disparate data sources ***but first you have to have data … <u>appropriate</u> data***

Data Requirements



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	cmpd	crystal_form	cell_volume	El1	El1_count	El2	El2_count	El3	El3_count	Z	cell_atom_count	cell_average_Z	r_min	r_max	r_average	mw	cell_weight	density_exp	e_density_ce
2	Y2SiO5	monoclinic	852.66	Y	2	Si	1	O	5	8	64	16.5	2.433	2.671	2.528	285.8942	2287.1536	4.454258788	1.23847723
3	Gd2SiO5	monoclinic																	
4	LuPO4	tetragonal																	
5	YPO4	tetragonal	286.53	Y	1	P	1	O	4	4	24	14.33333333	2.433	2.671	2.528				
6	Y3Al5O12	cubic	1734.92	Y	3	Al	5	O	12	8	160	13.9	2.433	2.671	2.528	593.61804	4748.94432	4.545415908	1.28190348
7	YAlO3	orthorhombic	201.95	Y	1	Al	1	O	3	4	20	15.2	2.433	2.671	2.528	163.885588	655.542352	5.390299556	1.505323
8	LuBO3-calcite	trigonal	339.21	Lu	1	B	1	O	3	6	30	20	2.433	2.671	2.528	233.7762	1402.6572	6.866554983	1.76881577
9	LuBO3-vaterite	hexago	104.92	Lu	1	B	1	O	3	2	10	20	2.433	2.671	2.528	233.7762	467.5524	7.399936828	1.90621425
10	BaF2	cubic	237.91	Ba	1	F	1	F	1	4	12	24.66666667	2.433	2.671	2.528	175.3238064	701.2952256	4.894904113	1.24416796
11	LaF3	trigonal	333.36	La	1	F	3			6	24	21	2.433	2.671	2.528	195.9007096	1175.404258	5.855038776	1.5118790
12	CaF2	cubic	163.78	Ca	1	F	2			4	12	12.66666667	2.433	2.671	2.528	78.0748064	312.2992256	3.166401948	0.92807424
13	Y2O2S	trigonal	82.1	Y	2	O	2	S	1	1	5	22	2.433	2.671	2.528	241.8755	241.8755	4.892205894	1.33982947
14	Y2O3	cubic	1191.62	Y	2	O	3			16	80	20.4	2.433	2.671	2.528	225.8099	3612.9584	5.034787912	1.36956412
15	InBO3	trigonal	310.83	In	1	B	1	O											0564617
16	ScBO3	trigonal	297.96	Sc	1	B	1	O											0684655
17	LuBO3-calcite	trigonal	339.21	Lu	1	B	1	O											6881577
18	LuBO3-vaterite	hexago	104.92	Lu	1	B	1	O	3	2	10	20	2.433	2.671	2.528	233.7762	467.5524	7.399936828	1.90621425
19	YBO3	monocli	326.73	Y	1	B	1	O	3	6	30	13.6	2.433	2.671	2.528	147.71505	886.2903	4.504462616	1.2487374
20	GdBO3	trigonal	1018.39	Gd	1	B	1	O	3	18	90	18.6	2.433	2.671	2.528	216.0592	3889.0656	6.341429173	1.643771C
21	LaPO4	monocli	307	La	1	P	1	O	4	4	24	17.33333333	2.433	2.671	2.528	233.876861	935.507444	5.060167299	1.3550488
22	ScPO4	tetragon	250.83	Sc	1	P	1	O	4	4	24	11.33333333	2.433	2.671	2.528	139.927271	559.709084	3.705432283	1.08439979
23	Y2SiO5	monocli	852.66	Y	2	Si	1	O	5	8	64	16.5	2.433	2.671	2.528	285.8942	2287.1536	4.454258788	1.23847723
24	Y3Al5O12	cubic	1734.92	Y	3	Al	5	O	12	8	160	13.9	2.433	2.671	2.528	593.61804	4748.94432	4.545415908	1.28190348
25	Y2O3	cubic	1191.62	Y	2	O	3			16	80	20.4	2.433	2.671	2.528	225.8099	3612.9584	5.034787912	1.36956412
26	ScBO3	trigonal	297.96	Sc	1	B	1	O	3	6	30	10	2.433	2.671	2.528	103.76511	622.59066	3.469769611	1.00684655
27	InBO3		310.83	In	1	B	1	O	3	6	30	15.6	2.433	2.671	2.528	173.6273	1041.7632	5.565473089	1.50564617
28	YBO3	mo	326.73	Y	1	B	1	O	3	6	30	13.6	2.433	2.671	2.528	147.71505	886.2903	4.504462616	1.2487374
29	ScPO4	tetra	250.83	Sc	1	P	1	O	4	4	24	11.33333333	2.433	2.671	2.528	139.927271	559.709084	3.705432283	1.08439979
30	YPO4	tetrago	286.53	Y	1	P	1	O	4	4	24	14.33333333	2.433	2.671	2.528	183.877211	735.		2005723E
31	LaPO4	monoclini	307	La	1	P	1	O	4	4	24	17.33333333	2.433	2.671	2.528	233.876861	935.		.3550488
32	Gd2O2S	trigonal	85.59	Gd	2	O	2	S	1	1	5	32	2.433	2.671	2.528	378.5638	37		3693772E
33	LaOCl		6.48	La	1	O	1	Cl	1	2	6	27.33333333	2.433	2.671	2.528	190.3579	380.7158	5.427571362	1.40796703
34	Gd2S2O		7.77	Gd	2	S	2	O	1	4	20	33.6	2.433	2.671	2.528	394.6294	1578.5176	6.589817794	1.68941850
35	Y2S2O		4.05	Y	2	S	2	O	1	4	20	23.6	2.433	2.671	2.528	257.9411	1031.7644	4.461169999	1.2290066

modified Balcerzyk | notes | cell_constants | sanderson charges | icsd distances | cleaning paren's | balcerzyk list

Ready

Sum=180

Basic Tasks of Materials Informatics

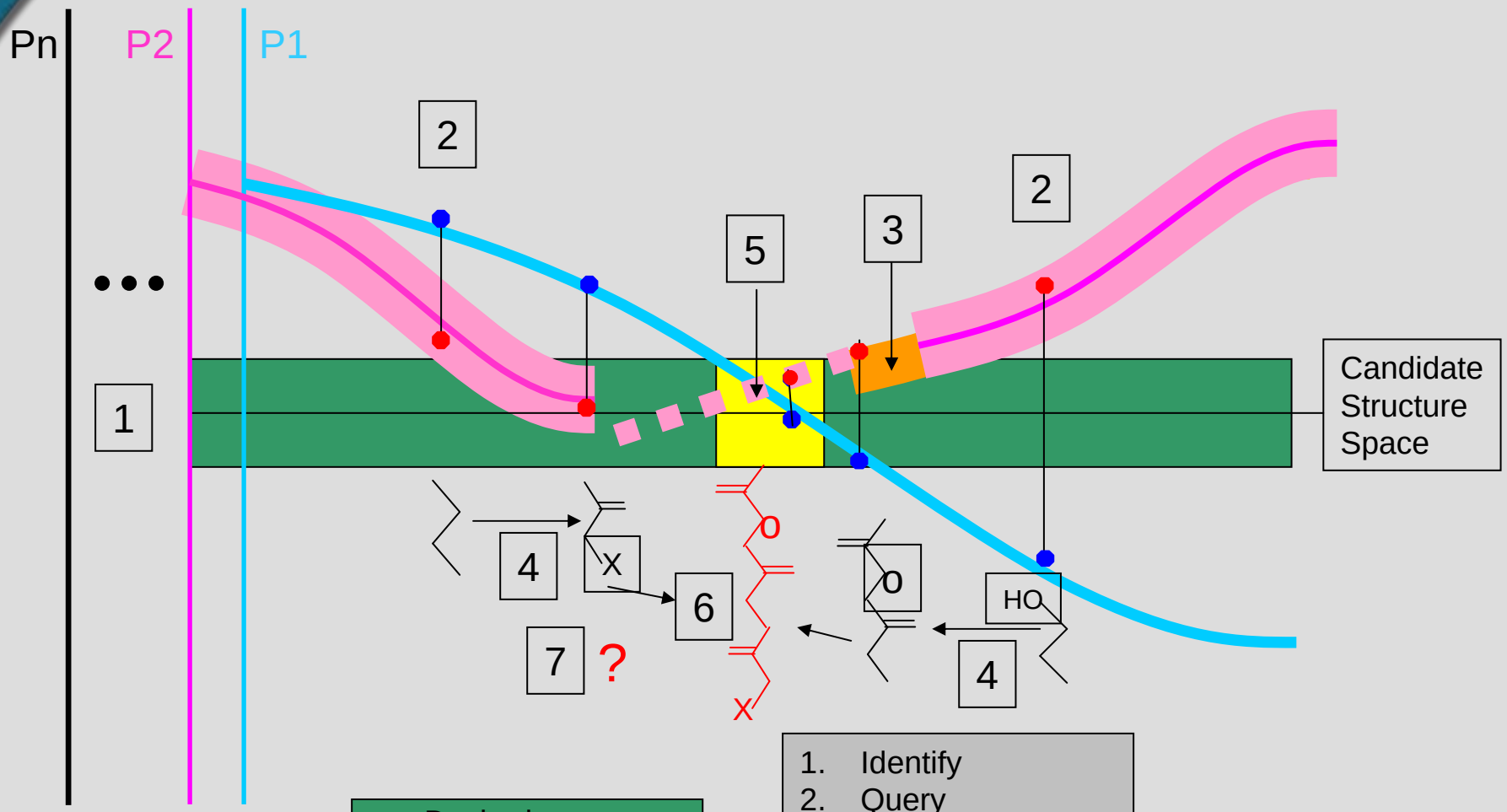
- Data Management – archival, anomoly detection
- Statistical – data transformation
- Information Hierarchy – joining multiscale information
- Classification – identification of rich/poor regions
- Regression – structure/property correlations
- Pattern Recognition
  - Diversity – needs to be appropriate to problem, multi-class
  - Feature development – mathematical relationships: maybe; design rules: yes.

Basic Requirement: Good Physical Model

Why Materials Informatics Matters

- Develops design rules
- Distinguishes a material's 'newness'
  - Precedented
  - Precedented in another application
  - Novel – not precedented
- Identifies 'candidate' materials based on multiple performance criteria
- Identifies trade-offs in materials specifications
- Manages database of problem-relevant materials and descriptors
- Finds anomalies in data – potential error, potential new science

Schematic for Multiple Property Mapping



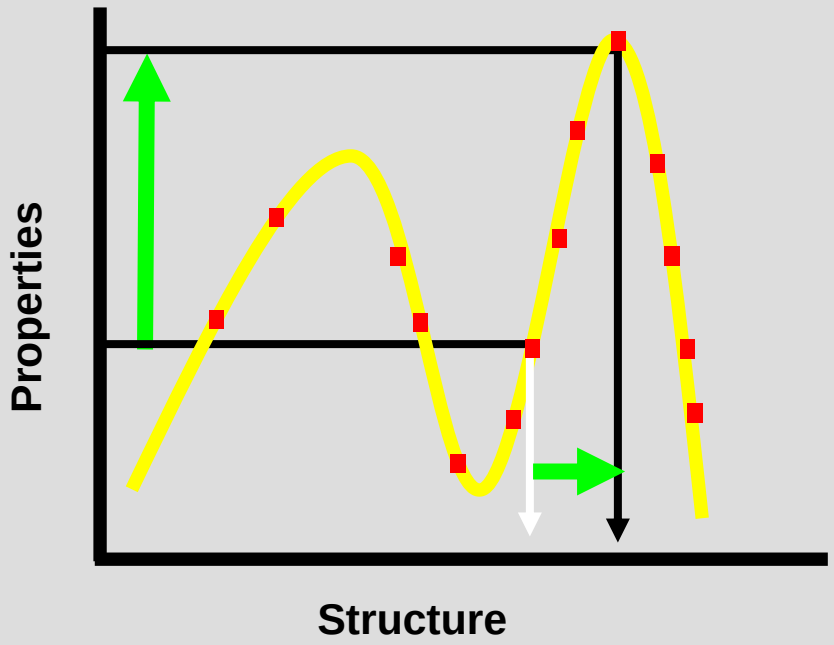
Desired range for P1, P2 (…Pn)

Hypothesized solution region

1. Identify
2. Query
3. Update models
4. Evolve
5. Extrapolate
6. Convert
7. Verify/Reject

Identify New Structures

Moving from existing to new materials



Goal

- Employ existing materials trends to propose new materials with improved properties

Tools

- Genetic and conventional evolution
- Structural signatures and models

Issues

- Incomplete or complex models

“Acme-Brand” Coating Candidates

Structural Repeat Unit	RI	CD	CAS #	Epoxide CAS #
-[C(X)(X)C(CX_3)(X)O]-	1.29	0.25	aaaaaa-bb-c	aaa-bb-c
-[C(X)(X)C(CX_3)(CX_3)O]-	1.29	0.25	-	aaa-bb-c
-[C(X)(X)C(CX_2CX_3)(X)O]-	1.29	0.24	aaaaaa-bb-c	aaaa-bb-c
-[C(CX_3)(CX_3)C(CX_3)(CX_3)O]-	1.29	0.21	-	aaaa-bb-c
-[C(X)(X)C(CX_2 H)(CX_3)O]-	1.30	0.35	-	aaaa-bb-c
-[C(CX_2 H)(CX_3)C(CX_2 H)(CX_3)O]-	1.30	0.33	-	-
-[C(X)(X)C(CX_2 H)(X)O]-	1.31	0.35	-	-
-[C(X)(X)C(X)(H)O]- *1	1.32	0.33		
-[C(OH)(CX_3)C(X)(X)]- *2	1.32	0.30		
-[C(OH)(X)C(X)(X)]- *2	1.33	0.33		
-[OC(X)(X)C(X)(X)OC(=O)]-	1.33	0.25		
-[C(X)(X)C(CX_3)(X)OC(=O)]-	1.34	0.25		
-[C(X)(X)C(CX_2 H)(X)OC(=O)]- *1	1.34	0.25		

* Designation denotes synthetically particularly difficult
1 β H results in HX elimination during synthesis
2 No obvious monomer reactive group for polymerization

Identifying Materials through Multi-stage Screening

Step	Rules applied	Possible Candidate Count
Basis	Binaries A_mB_n: m,n = 1,2,3 x 7 crystal systems	311,850
Element screen Function of element only F(A) or F(B)	Single crystal phase (Background Regulation) (High Cross-section) (Avoided Reactivity) (Avoided chemistry)	204,120 139,293 49,140 44,289 27,405
Formula screen Function of formula only F(AB)	(Valence Rules) (Crystal Polymorphs) Hygroscopic	2,240 24 n/a
Global screen Function of formula and structure	Band Gap Mechanical	n/a n/a

Structure Maps Revisited

- CALPHAD is a model-driven approach to the phase diagram of a material
- Are there other approaches to predicting the crystal structure of a material?
- Obviously, yes. → Structure map
  First principles calculation

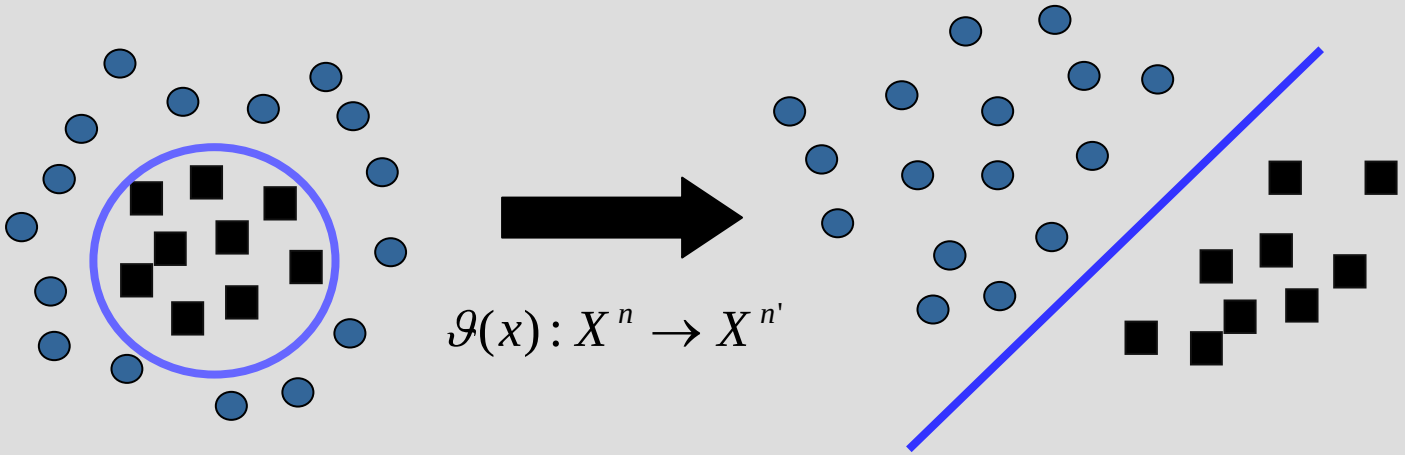
Basic Tasks for Materials Informatics

- Data Management
- Information Hierarchy
- Classification
- Regression
- Pattern Recognition
  - Diversity
  - Feature development

Mathematical Basis for Structure Plots

$$P_i(AB) \sim a_1(p_a) + a_2(p_b) + a_3(p_c) + a_4(p_d) \ldots$$

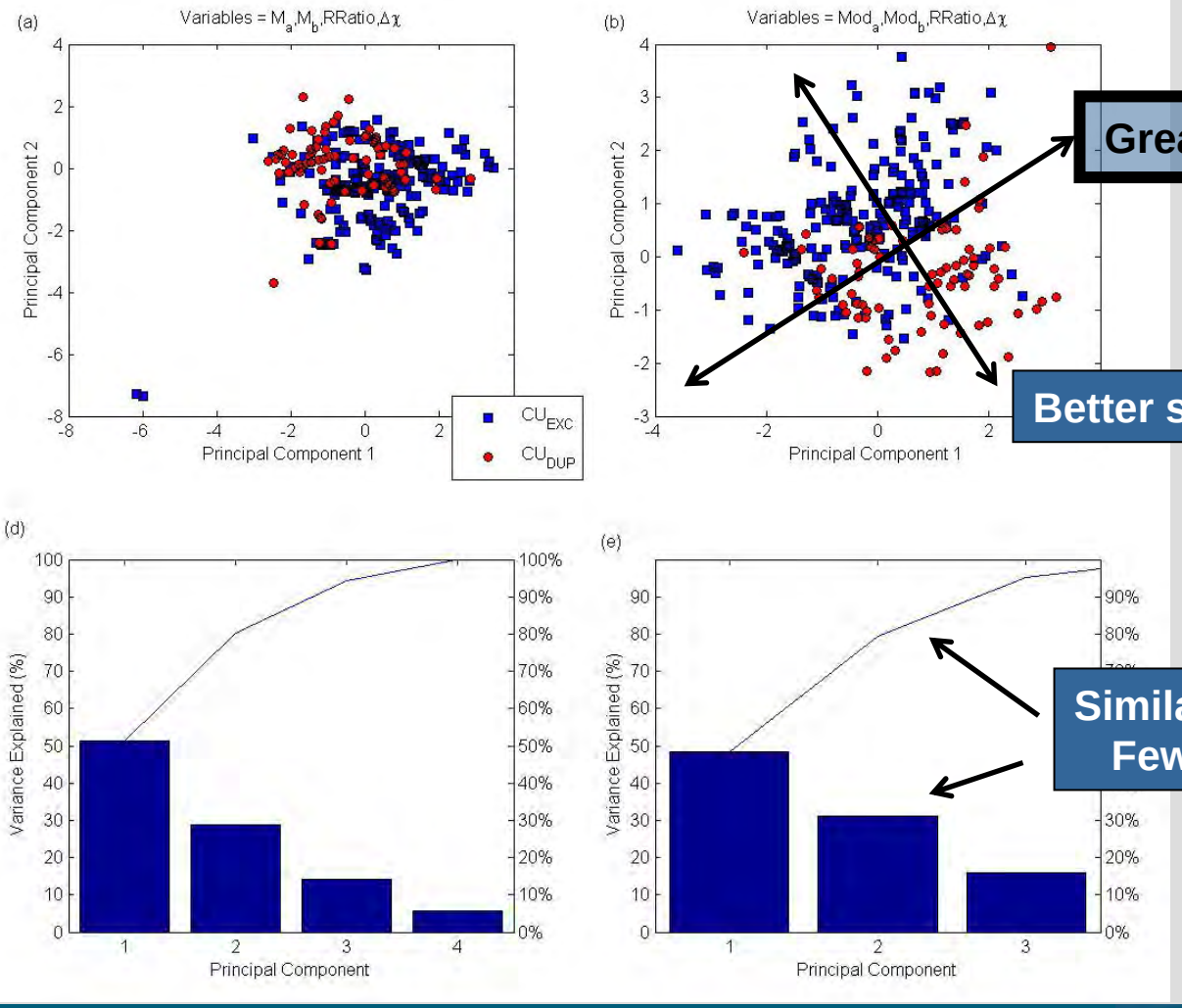
1. Variable selection (which measurements contribute to separability)
2. Reduced dimensionality representation (latent variables)



Model Development and Knowledge Extraction

Original Description

Revised Description

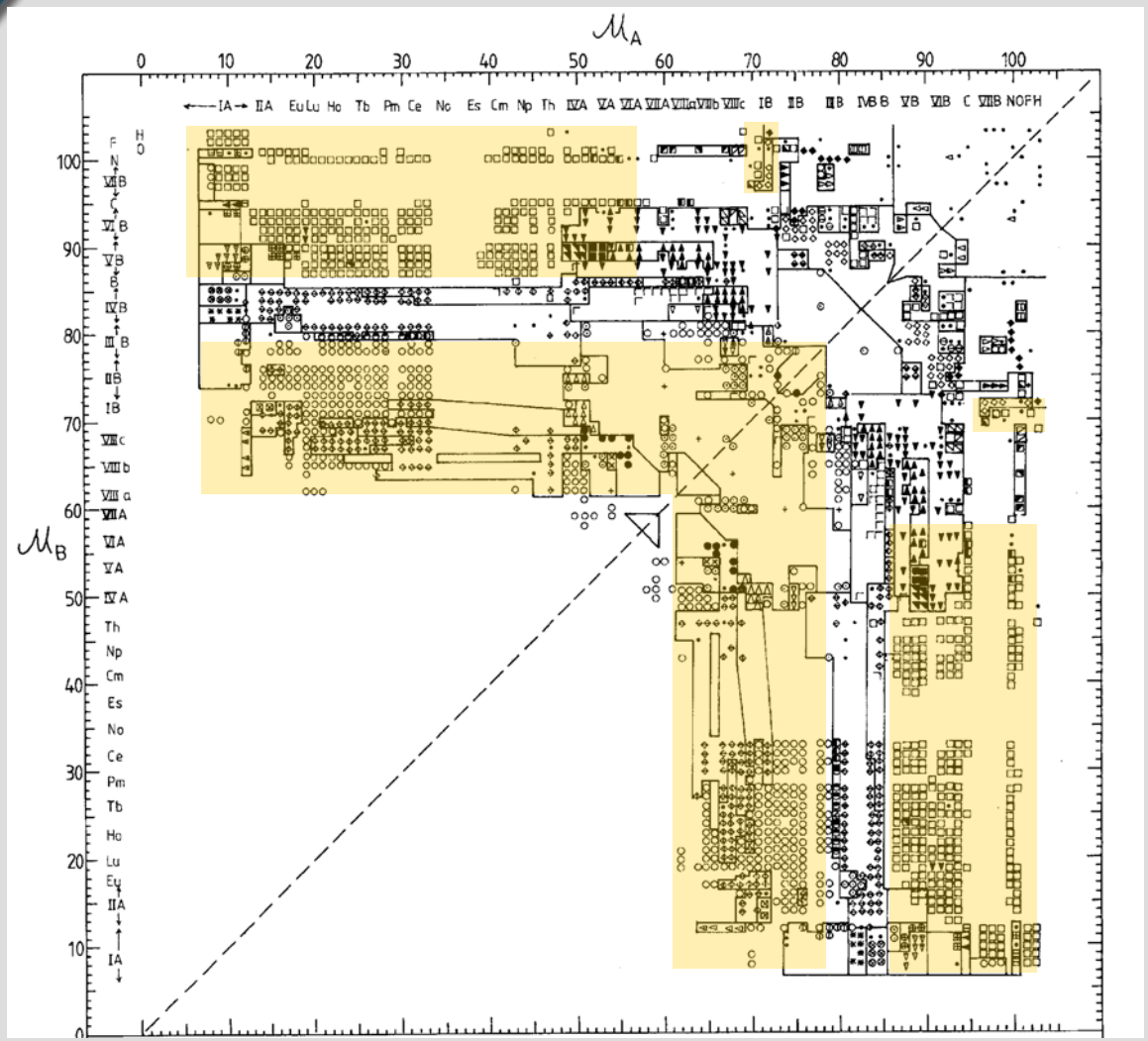


Greater variation

Better separation

Similar Description, Fewer Variables

Structure Maps Revisited: Pettifor



Cubic structure spaces

D.G. Pettifor, "The structures of binary compounds: Phenomenological structure maps," *J. Phys. C* 19 285-313 (1986).

Structure Maps Revisited: QSD/QFD

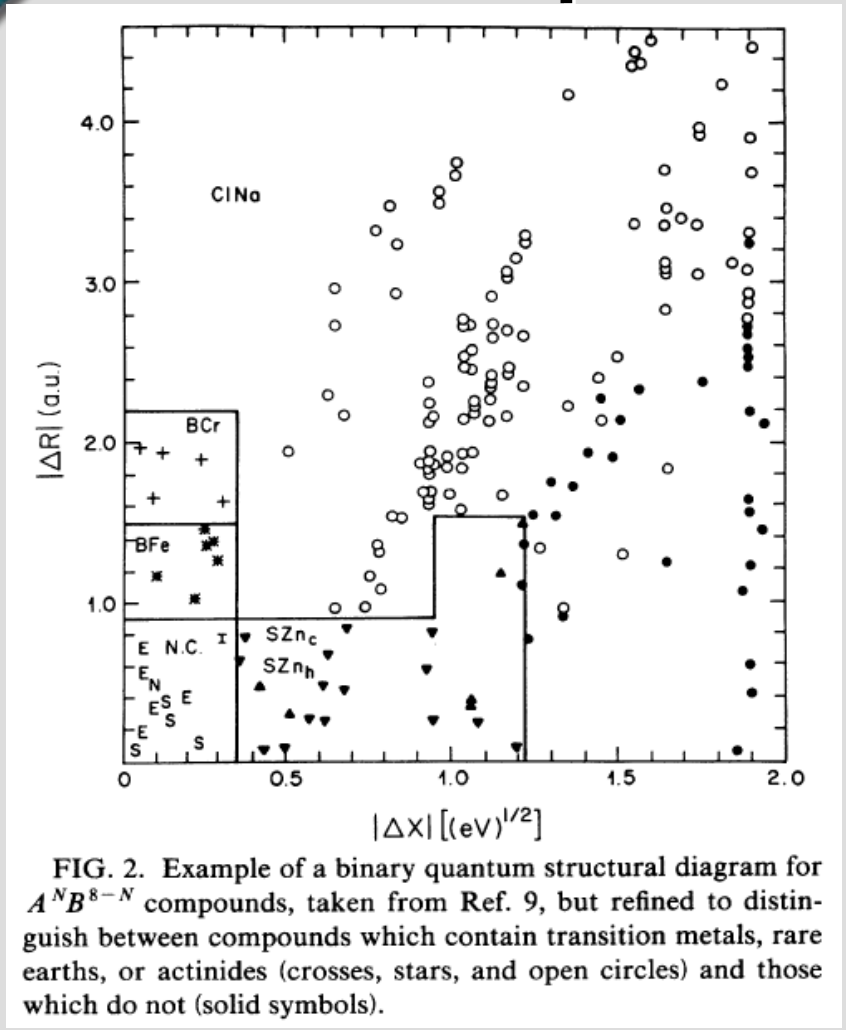


FIG. 2. Example of a binary quantum structural diagram for $A^N B^{8-N}$ compounds, taken from Ref. 9, but refined to distinguish between compounds which contain transition metals, rare earths, or actinides (crosses, stars, and open circles) and those which do not (solid symbols).

QSD: Quantum Structure Diagram

QFD: Quantum Formation Diagram

- 'Structure' concept extended to different properties
- Extensions to ternary, high-Tc, ferromagnetic, …

K.M. Rabe, J.C. Phillips, P. Villars, I.D. Brown, "Global multinary structural chemistry of stable quasicrystals, high-Tc ferroelectrics, and high-Tc superconductors," *Phys. Rev B* 45 7650-76 (1992).

Elemental Mappings: Pettifor

- 6 most distinct EPP groups
  - Atomic number
  - Group number
  - Mendeleev number
  - Cohesion energy
  - Electrochemical factor
  - Size
- Operations

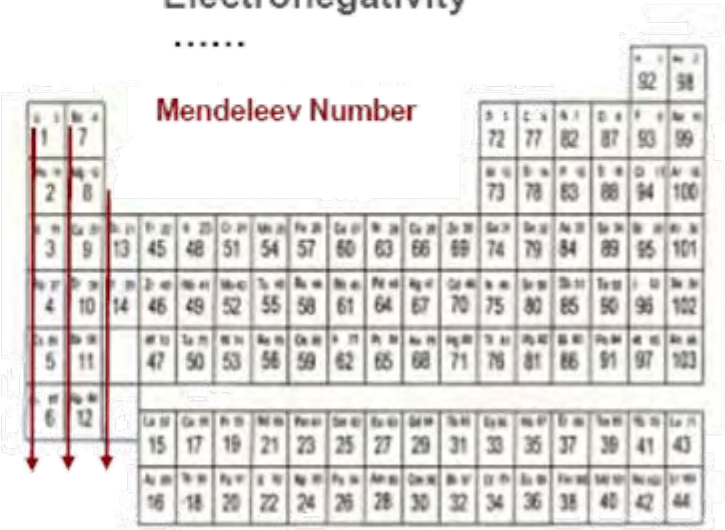
Sum	EP(A)+EP(B)
Difference	EP(A)-EP(B)
Product	EP(A)*EP(B)
Ratio	EP(A)/EP(B)
Maximum	Max(EP(A),EP(B))
Minimum	Min(EP(A),EP(B))

- EPP

  EP(tot) = EP(A) op EP(B)

Reference: www.umekkii.jp

- Example of EPs
  - Atomic number
  - Mendeleev Number
    (D. G. Pettifor, 1983, ...)
  - Atomic radii
  - Melting point
  - Electronegativity
  - .....



Signature / Model Development

Signature: *a (composite) descriptor that explains some data.*

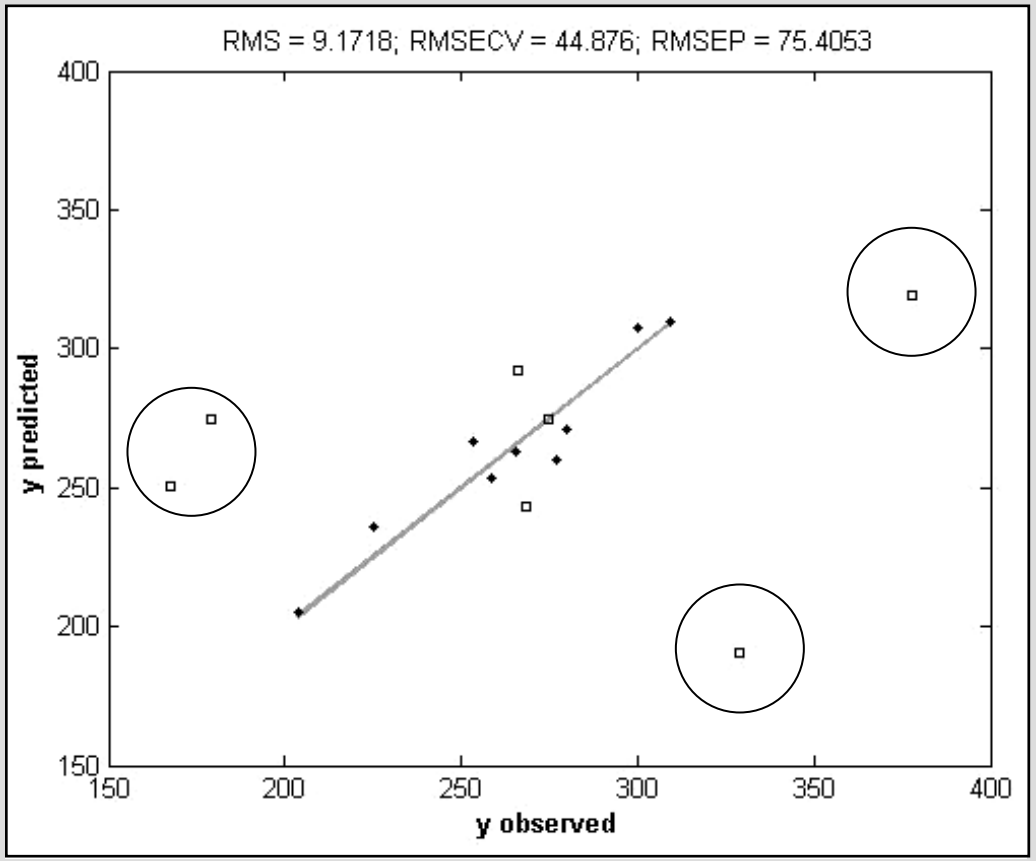
- Signatures/models:
  - improve knowledge of detector behavior
  - simplify and factor the overall design problem
- Tools
  - principal components analysis (PCA)
  - partial least squares (PLS)
  - projection pursuit (PP)
  - supervised learning theory
    - decision tree
    - neural net
    - support vector machine
- Issues
  - sparse data
  - incomplete descriptors

Model / Signature Examples

Model	Key signature
Protein function classifier	Sequence + property
Polymer glass-transition temperature	Effective sidechain/backbone mass ratio
Insecticide effectiveness	CLogP (Physical) Functional group (Chemical)
IR-spectrum determination of structure	1550 wavenumber peak → C=O …
Luminescence	Energy level differences between dopant and host

Need for Data: Model Development

Constructing a model for bulk modulus using elemental descriptors

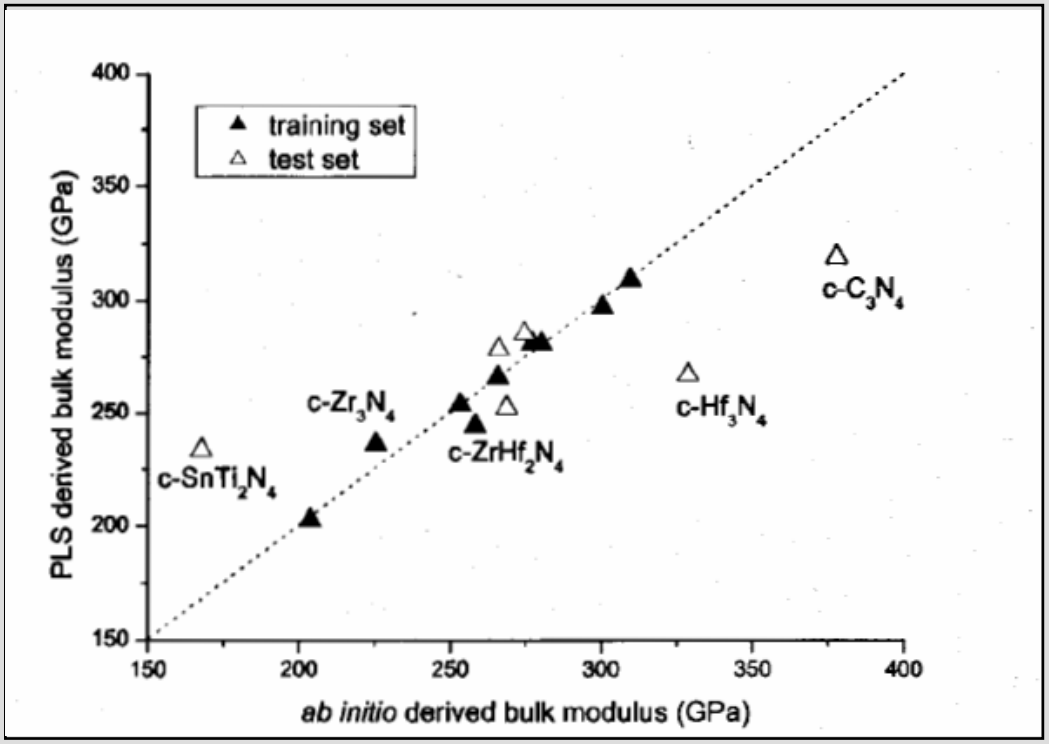


Descriptors
Topological
Mendeleev Number (A,B)
Ionization Potential
Electron Affinity
Non-iterative Charge
Molecular Weight
Cell Parameter

C. Suh and K. Rajan, "Combinatorial design of semiconductor chemistry for bandgap engineering: 'virtual' combinatorial experimentation," *Appl. Surf. Sci.* 223 148 (2004).

Better Data: Better Model Development

Constructing a model for bulk modulus using elemental descriptors



Descriptors
Average Electronegativity
Structural Parameter
Bond Length-AN
Bond Length-BN
Self-Consistent Charges

Reference: Suh, Rajan (2005); Ching, *J.Am.Cer.Soc.* 85 75-80 (2002).

Descriptors from First Principles Computations

- Descriptors for an AB compound fall into several classes:
  - Elemental $P_i(AB) \sim f(A_i,B_i)$
  - Compositional $P_i(AB) \sim f(AB_i)$
  - Structural $P_i(AB) \sim f(AB_i,CS_{AB})$

<u>Elemental</u>

- Size
- Heat
- Electrochemical
- Valence Electron
- Atomic Number
- Mendeleev Number
- Magnetic and electrical
- Thermodynamic

<u>Compositional</u>

- Thermodynamic; H, S, G
- Chemical Bonding
- Charge
- Structural
- Topological
- Electric and Magnetic Susceptibilities
- Mobilities
- Activation energies, kinetics

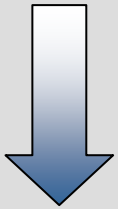
<u>Structural</u>

- Madulung Energies
- Dielectric Tensors
- Light scattering
- Topological

P. Villars, K. Brandenburg, M. Berndt, S. LeClair, A. Jackson, Y.-H. Pao, B. Igelnik, M. Oxley, B. Bakshi, P. Chen, S. Iwata, "Interplay of large materials databases, semi-empirical methods, neuro-computing, and first principles calculations for ternary compound former/nonformer prediction," *Eng. Appl. Art. Intell.* 13 497-505 (2000).

Limitations to the Data-Driven Approach

- Cannot provide insight into microscopic mechanism
- Limited experimental data leads to incomplete separation of different domains



Combining First Principles Approaches

Further Motivation for First Principles

Atomistic Level Simulation of Materials

- To explain electronic level and microscopic mechanisms for experimental measurements.
- To predict unknown substance and unknown properties.

First Principles Computations for Materials Informatics

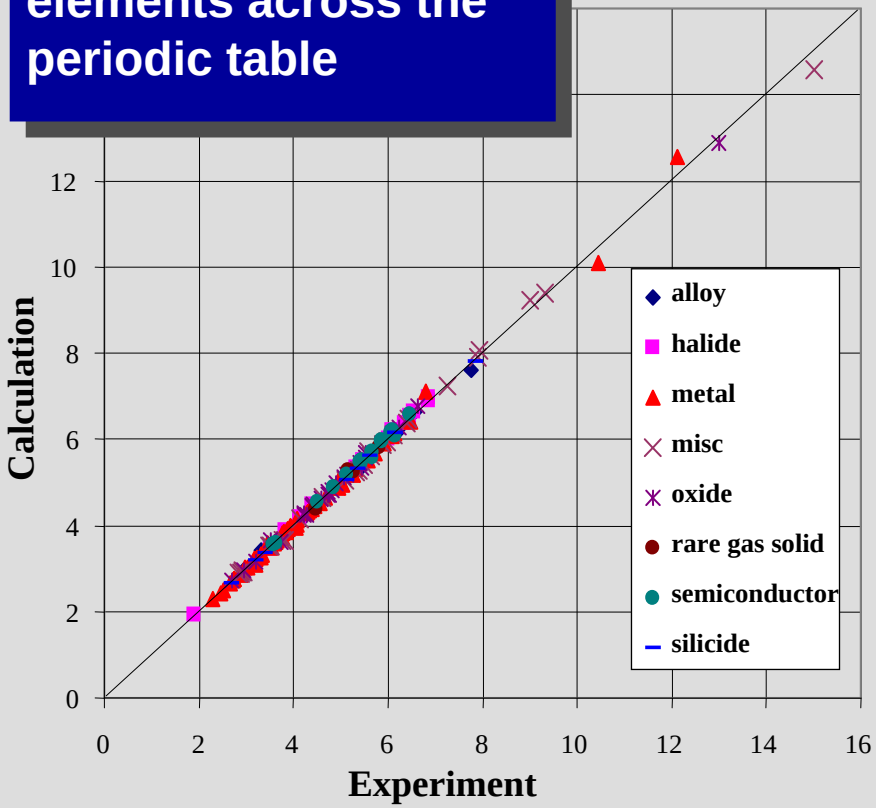
MRS Bulletin,

Sept. 2006 Issue

- Mechanical Properties and Structured Materials
- Catalysis and Surface Science
- Magnetism and Magnetic Materials
- Oxides and Minerals
- Semiconductors and Nanotechnology
- Biomaterials

Model Performance and Characteristics

181 lattice parameters for crystals containing elements across the periodic table

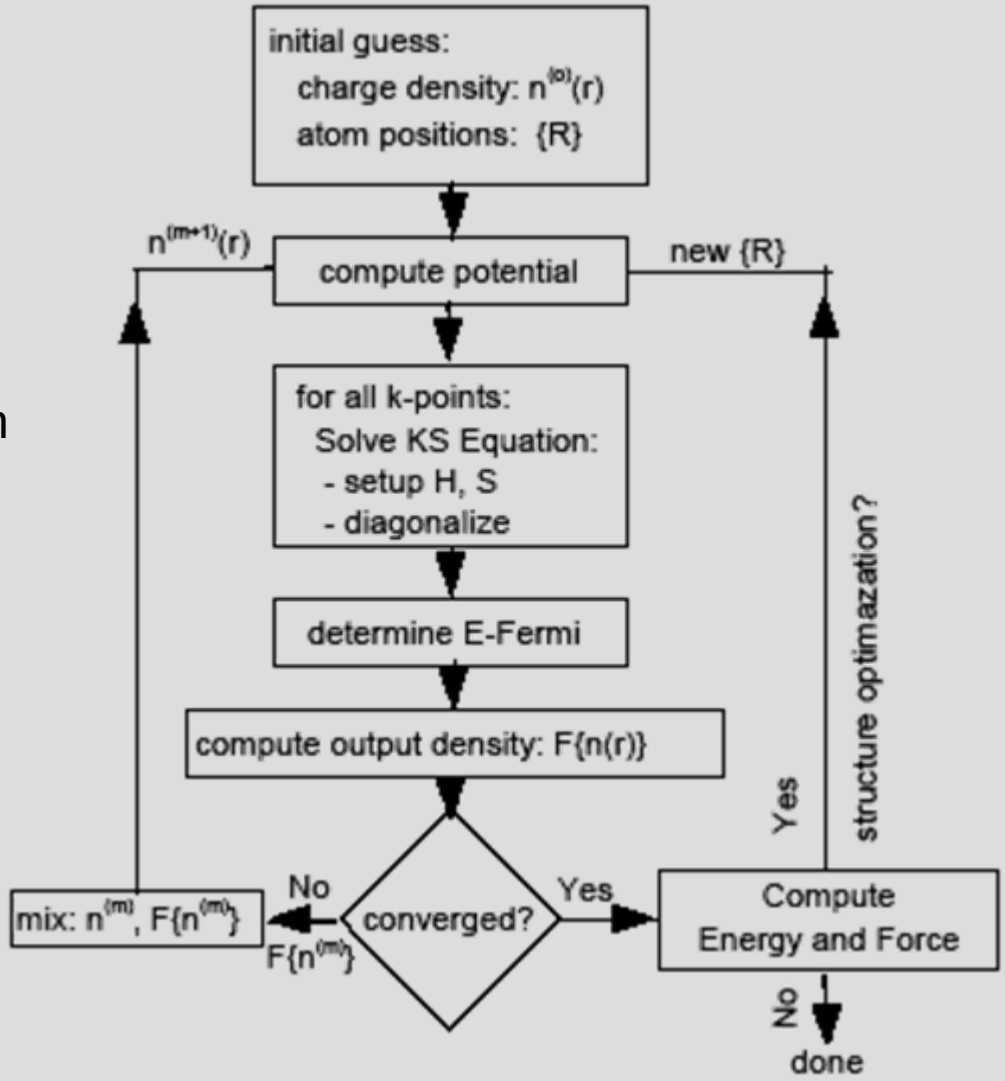


- Inputs: atomic numbers, crystal coordinates
- Energies, forces, stresses, structures and properties are ***PREDICTED*** with minimal input
- Most elements in the periodic table
- Molecules, solids, surfaces, and interfaces
- Validated through 1000's of publications in peer reviewed journals
- Computationally intensive

V. Milman, B. Winkler, J. A. White, C. J. Pickard, M. C. Payne, E. V. Akhmatskaya, R. H. Nobes, *Int. J. Quant. Chem.* 77, No5, 895-910 (2000).

First Principles Methodology

Typical loop structure of a first-principles code based on density functional theory as applied to solid state materials.



J. Grotendorst, S. Blugel, D. Marx. *Computational Nanoscience: Do It Yourself!* (Eds.), John von Neumann Institute for Computing, Julich, NIC Series, Vol. 31, ISBN 3-00-017350-1, pp. 85-129, 2006.

Descriptors from First Principles Computations

- Descriptors for an AB compound fall into several classes:
  - Elemental $P_i(AB) \sim f(A_i,B_i)$
  - Compositional $P_i(AB) \sim f(AB_i)$
  - Structural $P_i(AB) \sim f(AB_i,CS_{AB})$

<u>Elemental</u>

- Size
- Heat
- Electrochemical
- Valence Electron
- Atomic Number
- Mendeleev Number
- Magnetic and electrical
- Thermodynamic

<u>Compositional</u>

- Thermodynamic; H, S, G
- Chemical Bonding
- Charge
- Structural
- Topological
- Electric and Magnetic Susceptibilities
- Mobilities
- Activation energies, kinetics

<u>Structural</u>

- Madulung Energies
- Dielectric Tensors
- Light scattering
- Topological

Villars (2000)

Combining First Principles and Knowledge Base

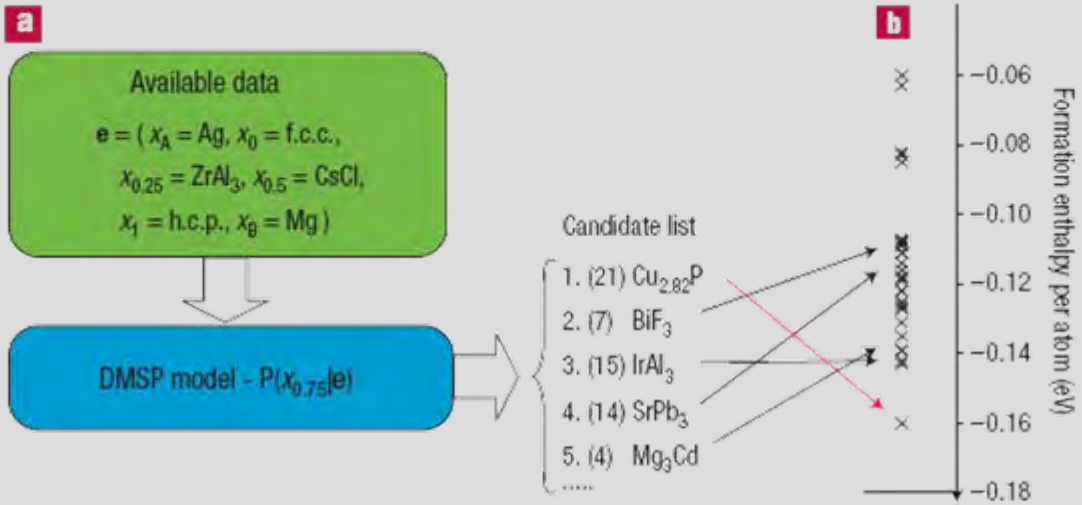


Figure 3 Predicting the structure of $AgMg_3$. a, DMSP prediction (candidate list) of the crystal structure of $AgMg_3$ on the basis of the limited data available at other compositions (green box). The structures are ordered by decreasing probability within the DMSP model. This ordering is compared with a ranking on the basis of the frequency with which these structures occur in the experimental database (parenthesized value in candidate list). **b**, *Ab initio* formation enthalpy (with respect to the pure elements) of the top five structures along with 26 additional structure types calculated to aid in verifying the prediction.

C.C. Fischer, K.J. Tibbetts, D. Morgan and G. Ceder, "Predicting crystal structure by merging data mining with quantum mechanics," *Nature Materials* 5 641-6 (2006) .

Quick Summary + First Principles

- Consistent with information-architecture
- “More” information, more diverse
- Added element for developing models and establishing their range of validity
- More than one valid model
  - alternate method for phase predictions
  - provides large basis of ‘chemical’ descriptors
  - computationally intensive
  - requires structural information
  - high accuracy