Informatics for Discovering the "Inorganic Gene" in Complex Inorganic Solids

Krishna Rajan,
Iowa State University

Outline

- Complexity and data dimensionality
- Mathematics of dimensionality reduction
- Crystallography…beyond 3 dimensions
- Extracting features from crystallographic information
- Impact: data driven modeling and simulation

Crystal Chemistry and Informatics

Mackay, Alan L. (2002) Generalized crystallography. *Structural Chemistry 13 (3-4) 215 – 220*

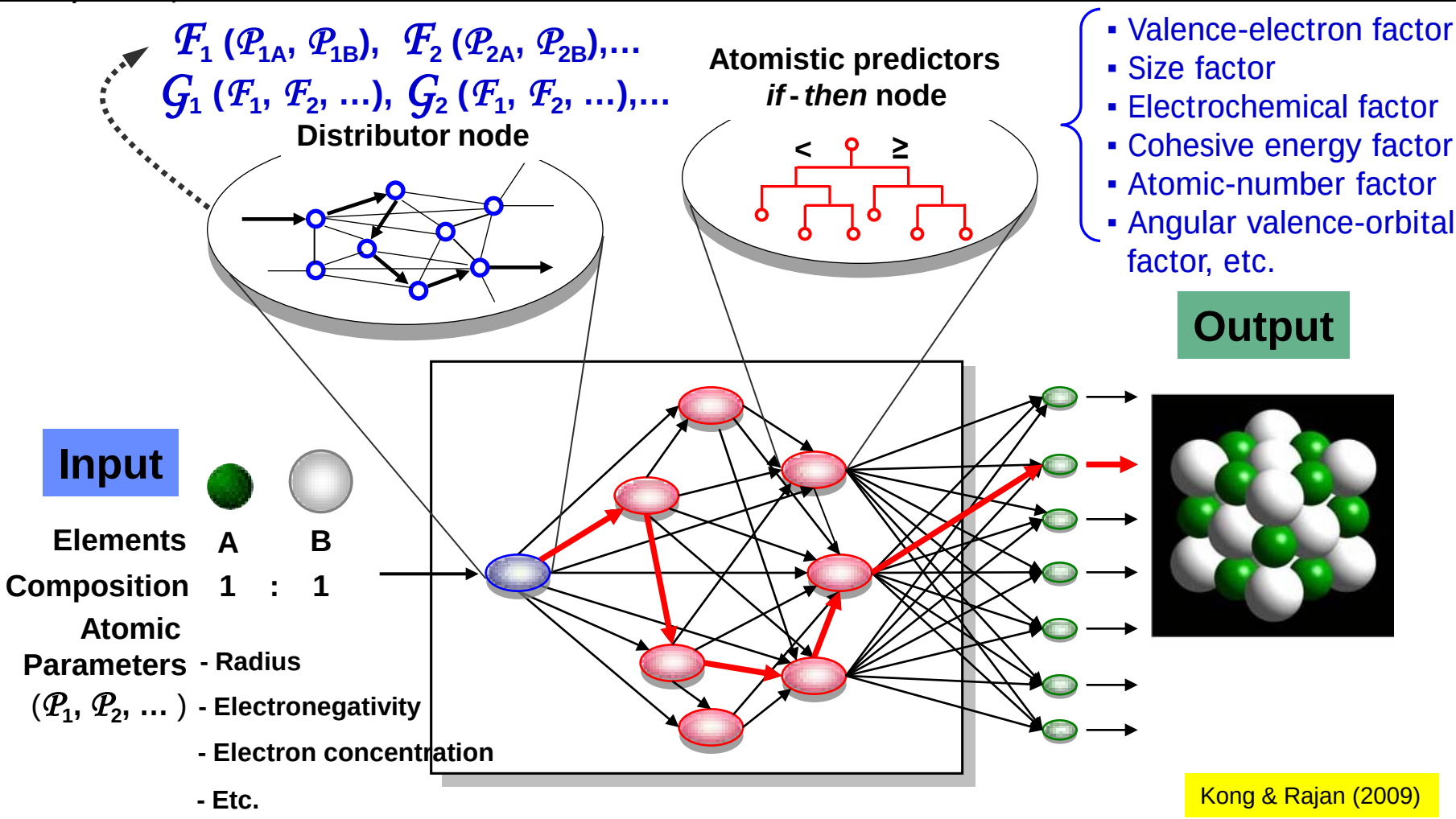
- *……It has been remarked that Linus Pauling, who formulated the principles of structural crystal chemistry, used only a perhaps a thousandth of the information now available, but nevertheless produced rules which have not required significant correction.*

- *The IUC have adopted a somewhat retrograde definition of crystal ……an undue emphasis to Fourier analysis and periodicity.*

- *……."a crystal is a structure the description of which is much smaller than the structure itself"……*

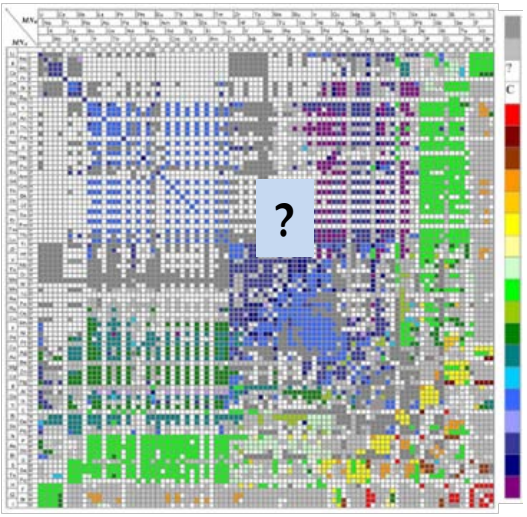
Information Carrier

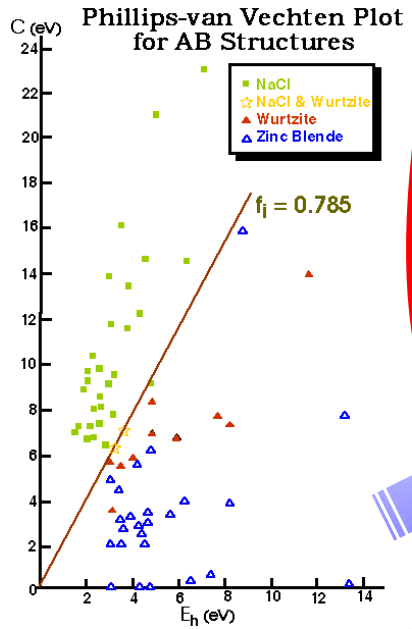
In a typical inorganic crystal the information about the structure of the crystal is carried in an non-localized form by the rules of interaction between the particular atoms present.
(MacKay,1986)

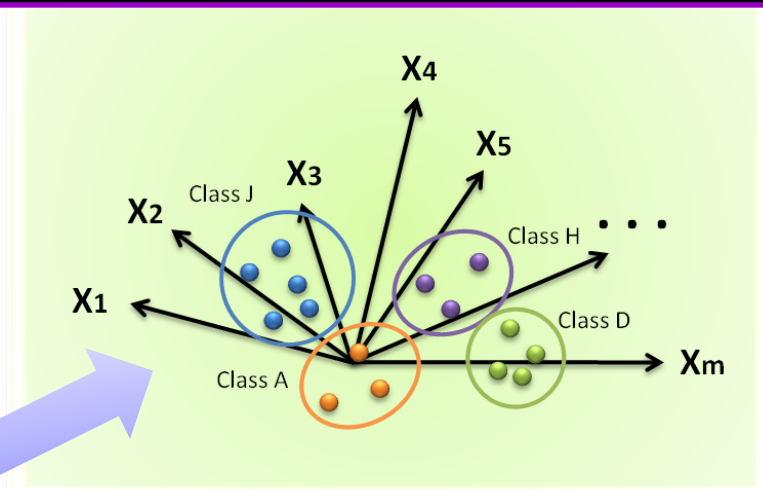


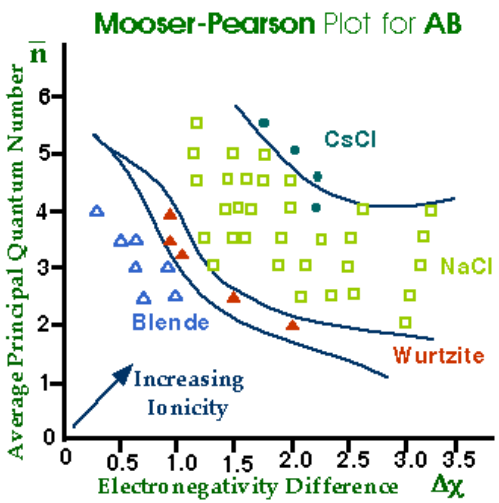
Kong & Rajan (2009)

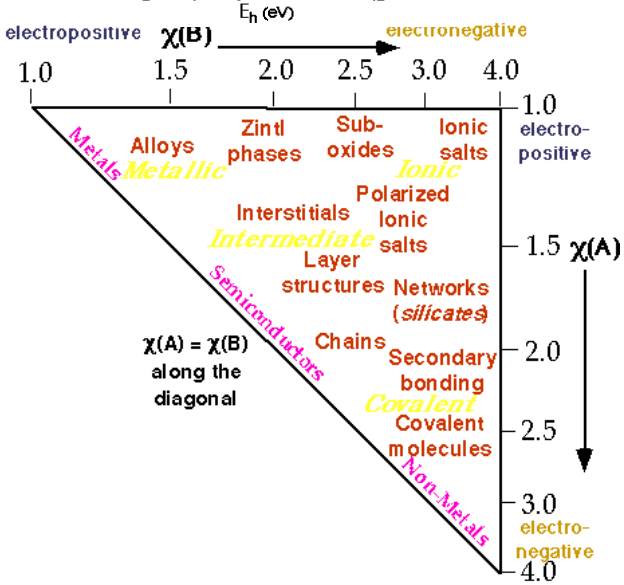
Rule Based Crystal Design: a high dimensional problem

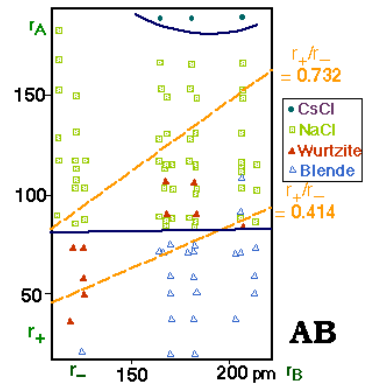


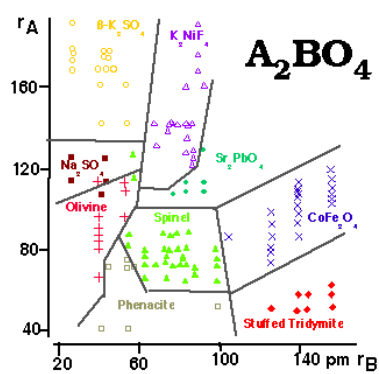












Computational Strategy

DATA INPUTS

•Crystallographic descriptors / semi-empirical electronic structure parameters / property data
•Descriptors for each element in compound…weighted by stoichiometry?

Reduce dimensionality : **PCA**

Rank & reduce descriptors: **loading plots &entropy calculations**

DATA MINING

Structure classification

•PCA bi –plots, SVM, Frequency Association Mining & Clustering analysis : **New structure maps/ classifications:**
•PCA + GA + NN hybrid techniques : **Classification structural distortions**

•**Establish design rules:**
•**Recursive partioning using Shannon entropy criterion for classification: "if-then rules" / Recursive partioning for prediction**
• PLS: **Discovery of new descriptors**

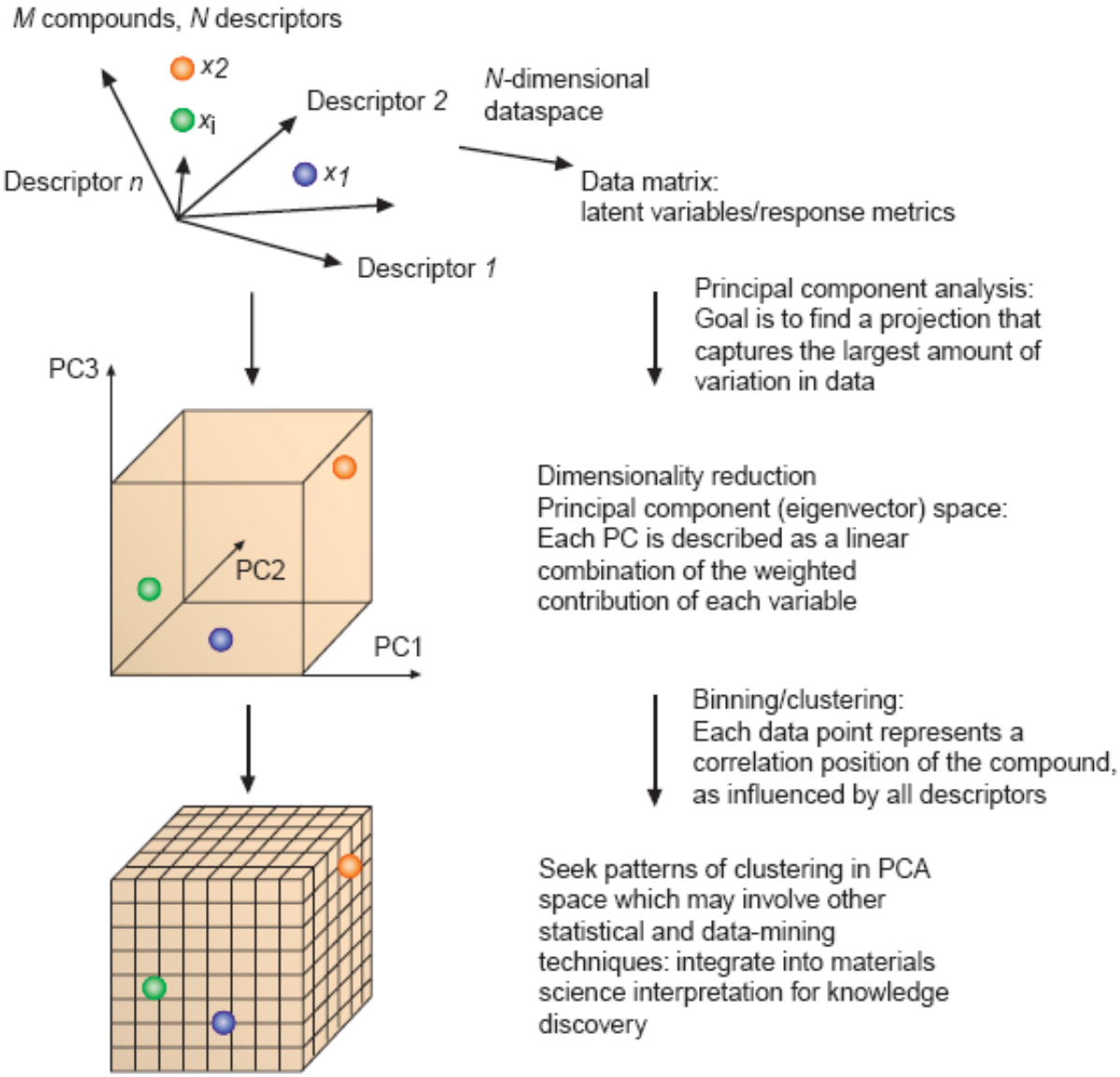
Structure-property prediction

• PLS & SVM- **QSAR for inorganic crystals**

Structure prediction:

•**Screen** clusters of compounds from above procedures- down select those for **energy** calculations

Dimensionality Reduction



Bajorath (2005)

Mathematical Basics

- Clustering: Given $\{X_1, \ldots, X_n\}$, build a function $f : \mathcal{X} \rightarrow \{1, \ldots, k\}$.
- Dimensionality reduction: Given $\{X_1, \ldots, X_n\} \in \mathbb{R}^D$, build a function $f : \mathbb{R}^D \rightarrow \mathbb{R}^d$
- Semi-supervised: Given $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ with $m << n$, build a function $f : \mathcal{X} \rightarrow \mathcal{Y}$
- Supervised: Given $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, build $f : \mathcal{X} \rightarrow \mathcal{Y}$

Connectivity and Complexity

Table 1. *Varieties of the connectivity matrix for N points*

$$[D]=\begin{bmatrix} d_{11}^2 & d_{12}^2 & . \\ d_{21}^2 & d_{22}^2 & . \\ . & . & . \end{bmatrix} \longleftrightarrow \begin{vmatrix} 0 & d_{12}^2 & d_{13}^2 \ldots 1 \\ d_{21}^2 & 0 & d_{23}^2 \ldots 1 \\ . & . & . \ \ldots 1 \\ 1 & 1 & 1 \ \ldots 0 \end{vmatrix} = (-1)N_2N-1[(N-1)!]^2V^2$$

d_{ij}^2 matrix $(N\times N)$

Determinantal expression for $(N-1)-$dimensional hypervolume. $(N+1)\times(N+1)$.

$\Updownarrow$

$$[g_{ij}]=\begin{bmatrix} \mathbf{a}_1^2 & \mathbf{a}_1^2 \,.\, \mathbf{a}_2 & \mathbf{a}_1 \,.\, \mathbf{a}_3 \ldots \\ \mathbf{a}_2 \,.\, \mathbf{a}_1 & \mathbf{a}_2^2 \ \ldots & \\ \ldots & \ldots & \ldots \end{bmatrix} \underset{\text{multiply}}{\overset{\text{factorize}}{\rightleftharpoons}} \begin{bmatrix} 0 & 0 & 0 & & & \\ x_1 & 0 & 0 & & & \\ x_2 & y_2 & 0 & & & \\ x_3 & y_3 & z_3 & 0 & & \\ x_4 & y_4 & z_4 & 0 & 0 & \\ x_5 & y_5 & z_5 & 0 & 0 & 0 \\ . & . & . & 0 & . & . \end{bmatrix}$$

Metric matrix $(N-1)\times(N-1)$

Matrix of orthogonal coordinates
Lower triangular matrix $N\times N$.

Various functions of d_{ij}

Invert

Invert

$[g_{ij}]$
Inverse metric matrix.

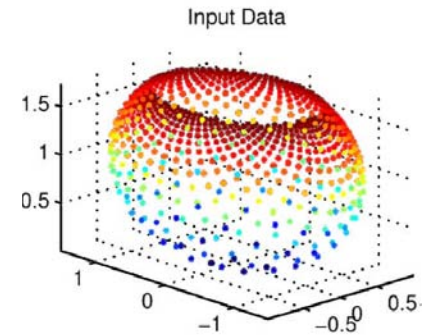
$$\begin{bmatrix} 0 & 1 & 1 & . & . \\ 1 & 0 & 1 & . & . \\ 1 & 1 & 0 & . & . \\ . & . & . & . & . \end{bmatrix} = [C]$$

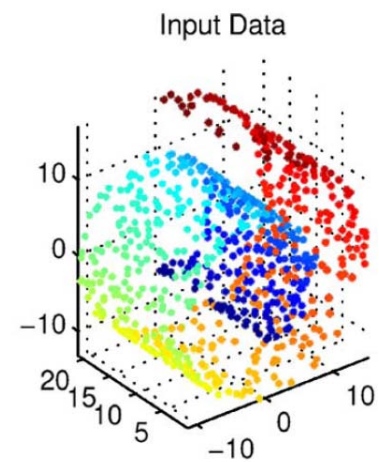
Adjacency matrix $(N\times N)$

Note:
$g_{ij}=\mathbf{a}_i \,.\, \mathbf{a}_j=\frac{1}{2}(d_{i0}^2+d_{j0}^2-d_{i9}^2)$
$d_{ij}^2=(\mathbf{a}_i-\mathbf{a}_j)^2=g_{ii}+g_{jj}-2g_{ij}$

$[\Omega_{ij}]$
Voronoi solid-angle matrix.

$[E_{ij}]$
Various expressions for energy from pairwise interactions.





Huang et.al. IEEE-Cybernetics- 2009

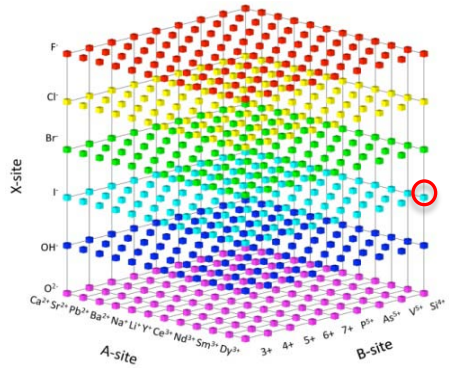
MacKay: Acta Cryst. (1974). A30, 440

Geometrical Complexity

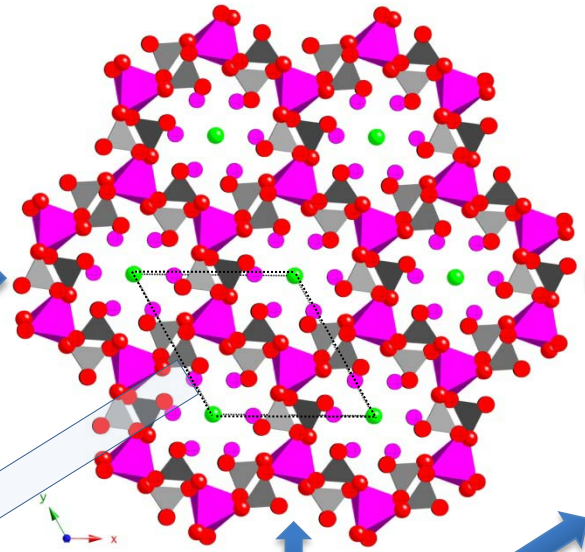
Balachandran & Rajan (2010)

Search Space

$A_{10}(BO_4)_6X_2$

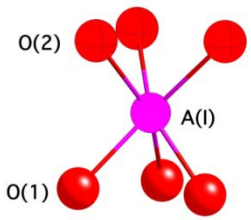


Apatite crystal structure *($P6_3/m$)*



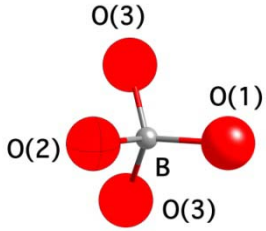
16 bond geometrical parameters were developed for describing one apatite

$d_{A(I)-O(1)}$ (Å)
$d_{A(I)-O(1)}^{A(I)z=0}$ (Å)
$\Delta_{A(I)-O}$ (Å)
$\Delta^{A(I)z=0}_{A(I)-O}$ (Å)
$\angle\psi_{A(I)-O(1)}$ (°)
$\angle\psi^{A(I)Z=0}_{A(I)-O(1)}$ (°)
$\angle\delta_{A(I)}$ (°)
$\angle\varphi_{A(I)}$ (°)
$\angle\alpha_{A(I)}$ (°)



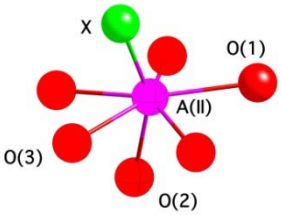
$A(I)O_6$ polyhedra

$d_{<B-O>}$ (Å)
$\angle<\tau_{O-B-O}>$ (°)



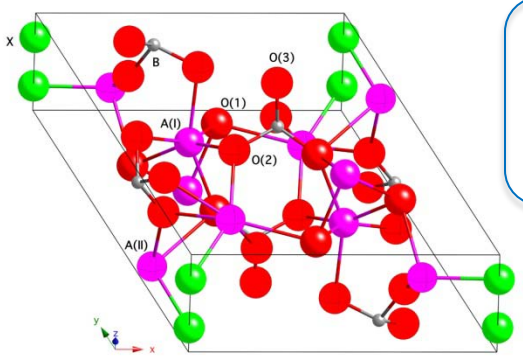
BO_4 tetrahedra

$\rho_{A(II)}$(Å)
$d_{A(II)-X}$ (Å)
$\angle\alpha_{A(II)}$ (°)
$d_{A(II)-O(3)}$ (Å)
$\angle\phi_{O(3)-A(II)-O(3)}$ (°)



$A(II)O_6X_{1,2}$ polyhedra

Unit cell of apatite



Spatial distribution of atoms

Atomic coordinates

$A(I)_z$, $A(II)_x$, $A(II)_y$, B_x, B_y, $O(1)_x$, $O(1)_y$, $O(2)_x$, $O(2)_y$, $O(3)_x$, $O(3)_y$, $O(3)_z$, X_z

Minimum energy solution

*# of atoms per apatite unit cell – **42***
*# of independent bond geometrical parameters – **16***
*# of chemical/physical descriptors for each element – **25***
*# of possible low symmetry modifications – **5 space groups***

Information Partioning

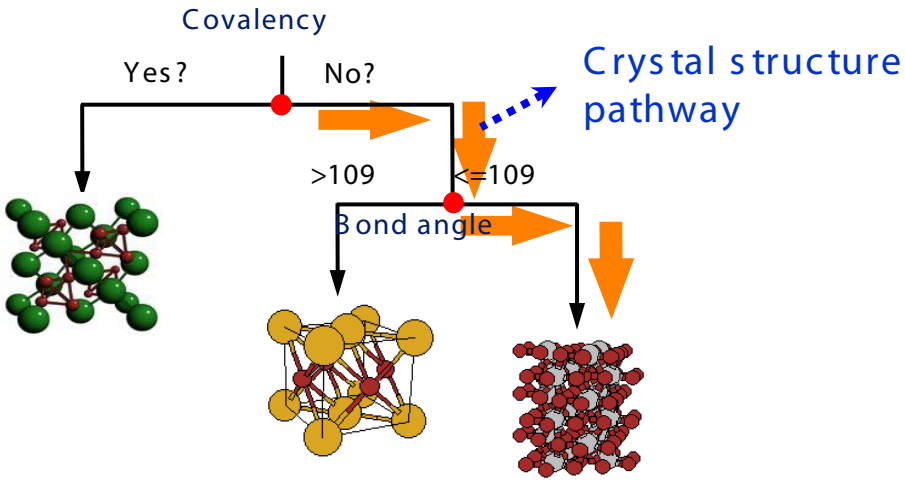
- Each compound in a AB_2 database, $S\{s_1, s_2, \ldots, s_k\}$, is characterized by l parameters, $V = V(v_1, v_2, \ldots, v_l)$, V being the parameter space, are assigned to one of the classes, c_i $C = \{c_1, c_2, \ldots, c_m\}$ (i.e., structure types), where i {1, …, m}.

- A classification $\Phi = \{\Phi_1, \Phi_2, \ldots, \Phi_n\}$ is constructed via recursive partioning of high dimensional data with the objective is to make the minimum number of dissimilar classes in each partitioned subspace.

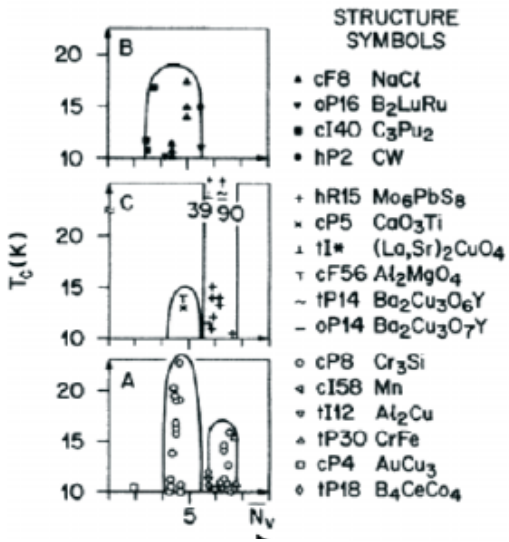
- Partition-based classification is carried out to sort out different crystal structures, where information entropy serves as a splitting criterion. The classification model is built based on the information of the probability function achieved using existing data set.

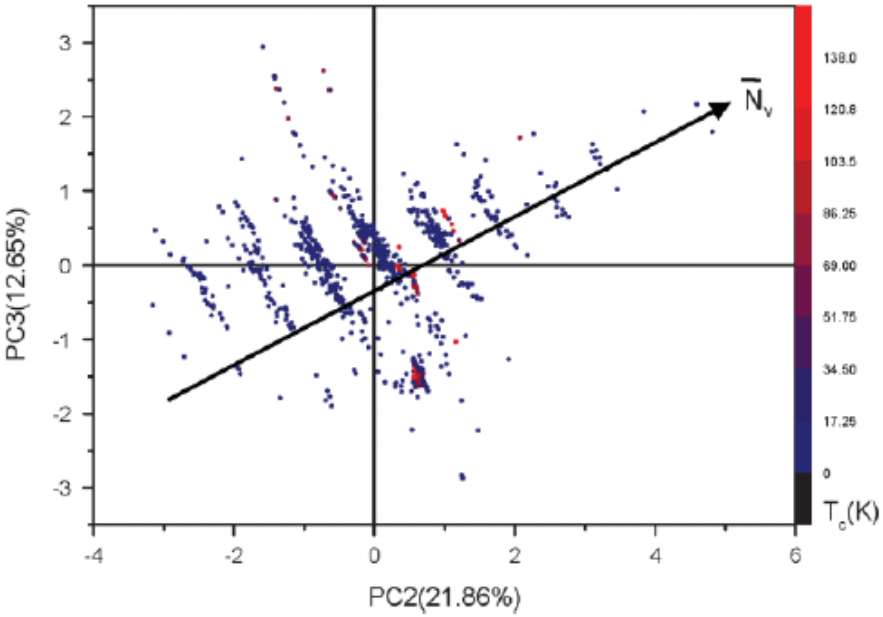
- Distribution of compounds in m-dimensional parameter space is described by means of the probability, $p(c_i | t)$, the frequency of the occurrence of a class, c_i at a subspace t

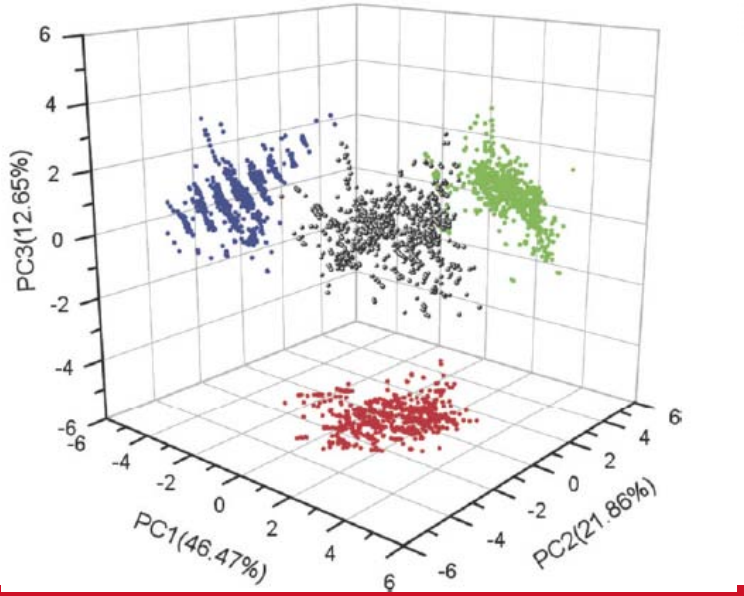
- Resulting classification as a mapping process (M), M: S→C, uncovers patterns in the data that can be developed as sets of useful rules



Clustering and Patterns



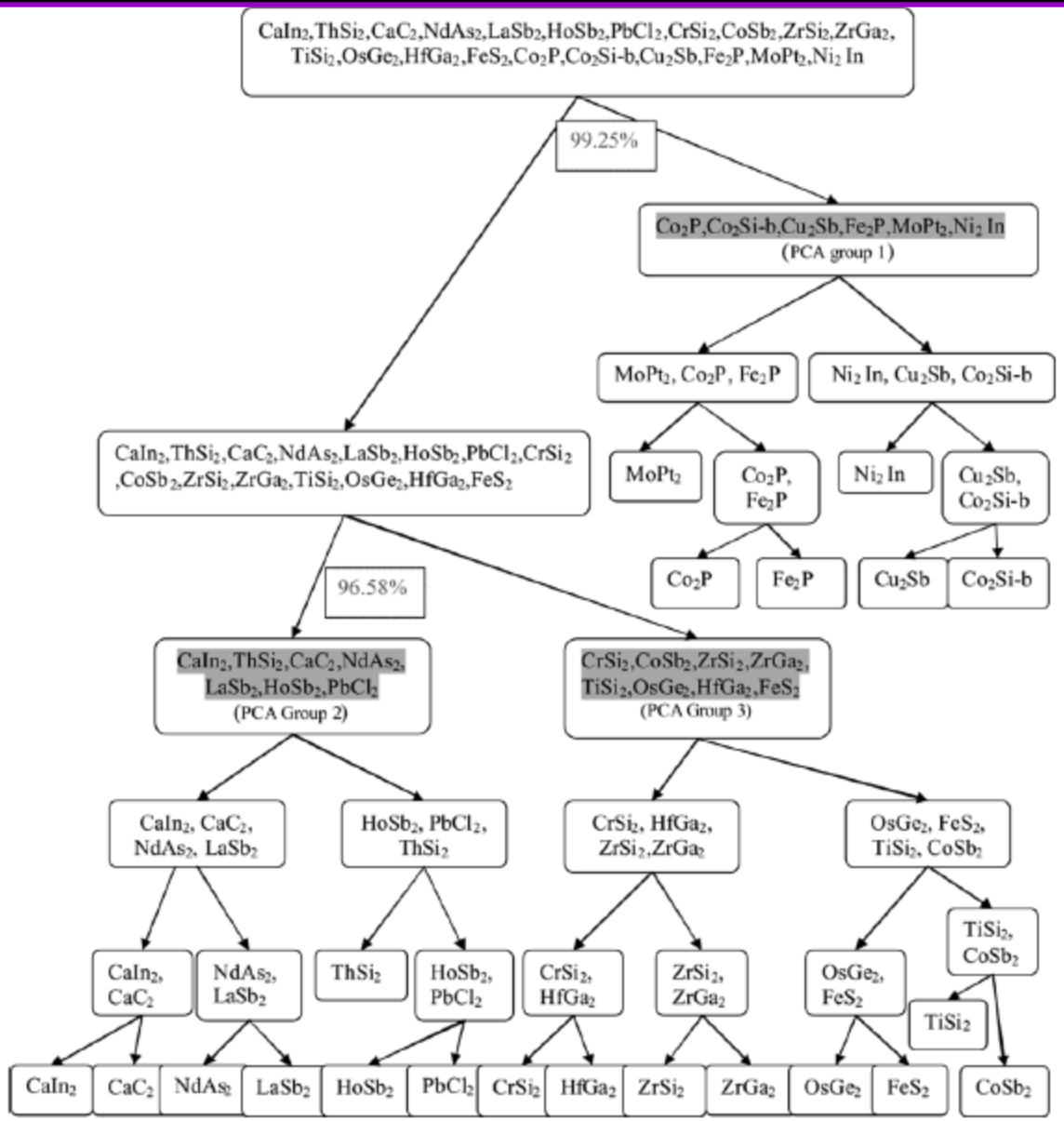




K. Rajan et al.: Materials Informatics Tools

Statistical Analysis and Data Mining, Vol. 1 (2009)

Defect Detection



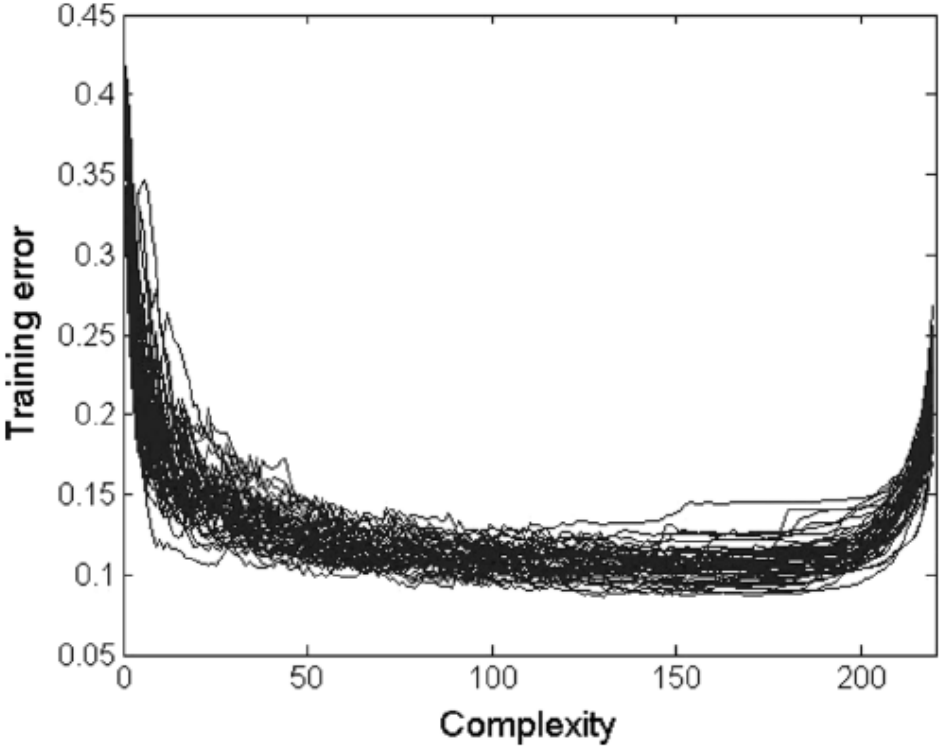
Agarwal et. al
Mater. Manuf. Proc. (**24** 2001-8 2009)

Complexity – Uncertainty Tradeoff

$A_{8a} - \leq_{16c}$, $B_{16d} - \Delta_{8b,48f}$
$A_{8a} - \Delta_{48f}$
$A_{8a} - \Delta_{48f}^{2ndn.n.}$, $B_{16d} - B_{16d}$, $B_{16d} - \leq_{16c}$
$A_{8a} - B_{16d}$, $B_{16d} - \Delta_{8}$, $B_{16d} - \Delta_{48f}^{2ndn.n.}$
$A_{8a} - A_{8a}$, $A_{8a} - \Delta_{8b}$
$(N_{32e} - N_{32e})_1$ ("unshared" anions)
$(N_{32e} - N_{32e})_2$ ("shared" anions)
$(N_{32e} - N_{32e})_3$
$\Delta_{8b} - N_{32e}$
$(\Delta_{48f} - N_{32e})_1$
$(\Delta_{48f} - N_{32e})_2$
$\leq_{16c} - N_{32e}$
$BL_{A\text{-}N}/BL_{B\text{-}N}$ (or A_{8a}-N_{32e}/B_{16d}-N_{32e})
Bond order between A (or B) and N (BO_{AN}, BO_{BN})
Relative stabilization energy per formula unit (ΔE)**
Burdett's parameters based on Zunger's total sizes of the effective core of atoms (r_σ^A and r_σ^B) [10, 11]
Zunger's parameterization scheme for AB compound (for nitrides, R_σ^{AN} and R_σ^{BN}) [

$\angle$A-N-A
$\angle$N-B-N
$\angle$B-N-B
$\angle$N-A-N
$\angle$A-N-B

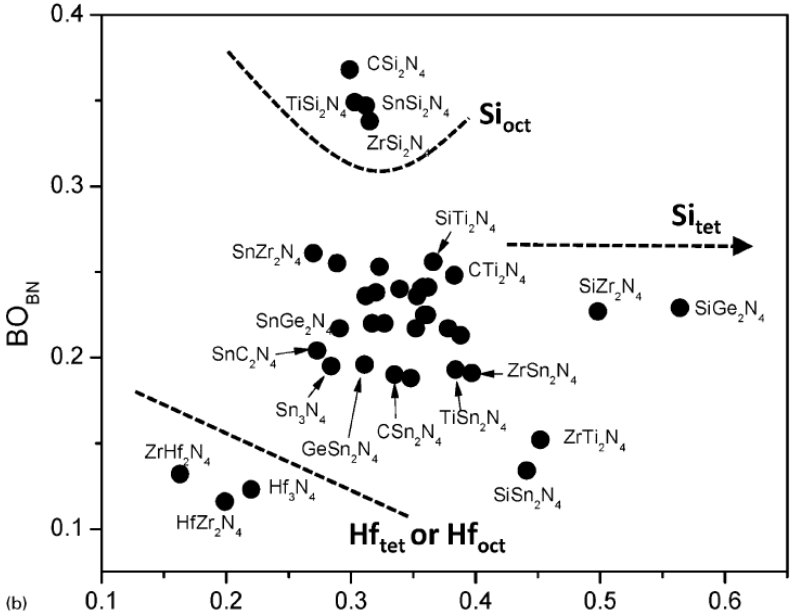
$V_{tet}^{A_{8a}}$
$V_{tet}^{A_{8b}}$
$V_{tet}^{A_{48f}}$
$V_{oct}^{B_{16d}}$
$V_{oct}^{B_{16e}}$
$V_{oct}^{N_{32e}}$

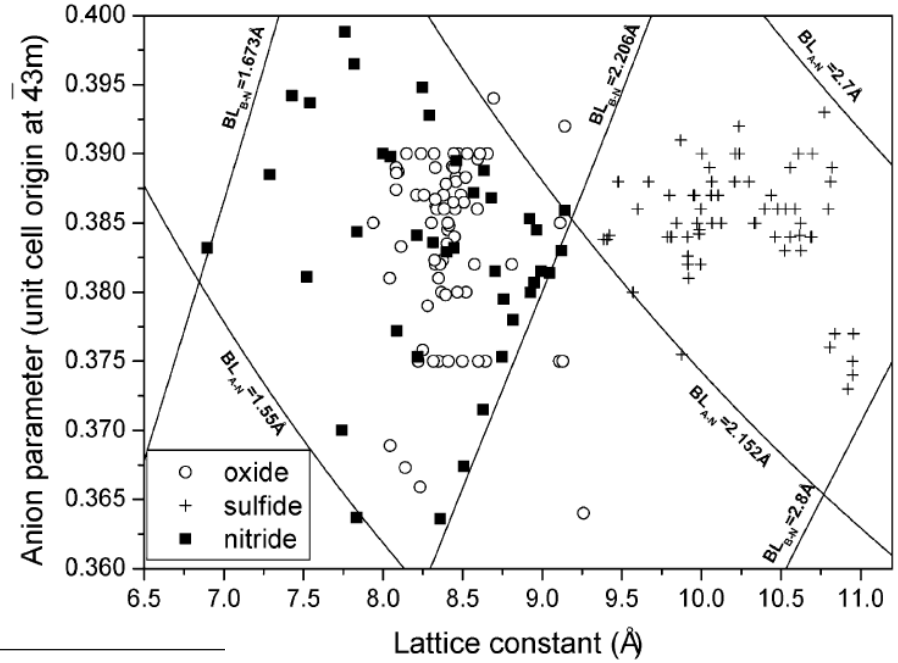


Lattice constant (a)
Anion parameter (u)
A-N bond length ($BL_{A\text{-}N} = A_{8a} - N_{32e}$)*
B-N bond length ($BL_{B\text{-}N} = B_{16d} - N_{32e}$)*
Effective charge on each atom (Q_{tet}^*, Q_{oct}^*, Q_N^*)
Total crystal bond order (BO_{Cry})

Petterson et.al.
Mater. Manuf. Proc. (2009)

Discovering New Complexities

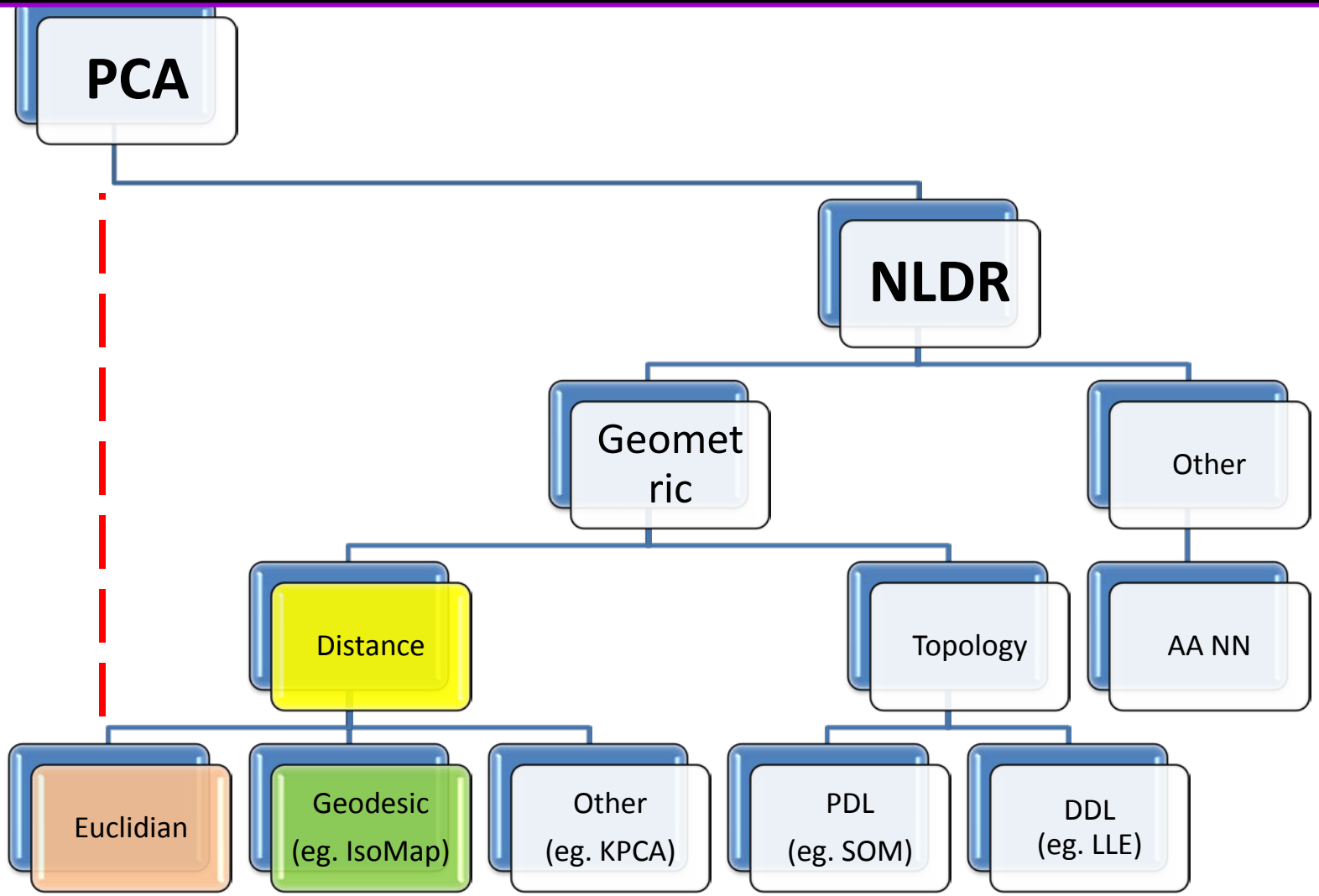




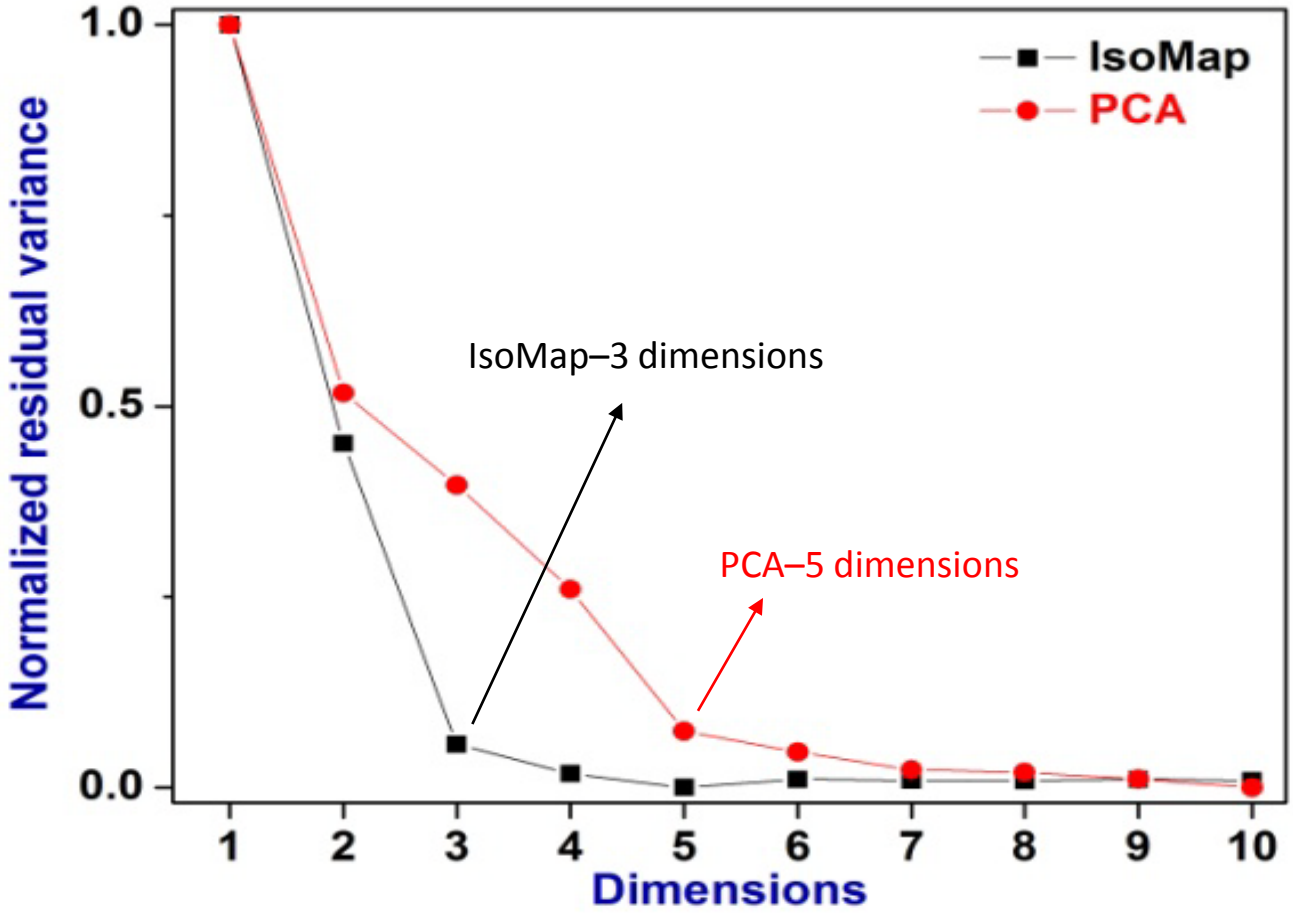
Origin of descriptors	Descriptors and its ID
Ab initio derived parameters by Ching *et al.*[4]	1. a (lattice constant), 2. u (anion parameter), 3. $BL_{A-N}=A_{8a}-N_{32e}$, 4. $BL_{B-N}=B_{16d}-N_{32e}$, 5. Q^*_A (effective charge), 6. Q^*_B, 7. Q^*_N, 8. BO_{Cry} (bond order of crystal), 9. BO_{AN}, 10. BO_{BN}
Pseudopotential orbital radii related parameters[24–26,28,29,32,33]	11. r^A_σ, 12. r^B_σ13. R^{AN}_σ, 14. R^{BN}_σ, 15. R^{AN}_π, 16. R^{BN}_π, 17. $\overline{R_\sigma}$, 18. $\overline{R_\pi}$, 19. PR (weighted differences of Zunger's pseudopotential radii sums)
Crystallographic parameters (bond lengths and polyhedral volumes) from a and u	20. $A_{8a}-\square_{16c}$ or $B_{16d}-\Delta_{8b,48f}$, 21. $A_{8a}-\Delta_{48f}$, 22. $A_{8a}-\Delta^{2nd\ n.n.}_{48f}$, $B_{16d}-B_{16d}$, or $B_{16d}-\square_{16c}$, 23. $A_{8a}-B_{16d}$, $B_{16d}-\Delta_{8b}$, or $B_{16d}-\Delta^{2nd\ n.n.}_{48f}$ 24. $A_{8a}-A_{8a}$ or $A_{8a}-\Delta_{8b}$, 25. $(N_{32e}-N_{32e})_1$, 26. $(N_{32e}-N_{32e})_2$, 27. $(N_{32e}-N_{32e})_3$, 28. $\Delta_{8b}-N_{32e}$, 29. $(\Delta_{48f}-N_{32e})_1$, 30. $(\Delta_{48f}-N_{32e})_2$, 31. $\square_{16c}-N_{32e}$, 32. BL_{A-N}/BL_{B-N}, 33. $V^{(A-8a)}_{tet}$, 34. $V^{(A-8b)}_{tet}$, 35. $V^{(A-48f)}_{tet}$, 36. $V^{(B-16d)}_{oct}$, 37. $V^{(B-16e)}_{oct}$, 38. $V^{(N-32e)}_{oct}$
Interbond angles from spinel structure	39. $tf == "P7DC6"char47$A–N–A, 40. $tf == "P7DC6"char47$N–B–N, 41. $tf == "P7DC6"char47$B–N–B, 42. $tf == "P7DC6"char47$A–N–B, †. $tf == "P7DC6"char47$N–A–N (constant=109·471°)
Derived parameters (combinations of parameters or miscellaneous)[23]	43. $Q^*_N-Q^*_B$, 44. $Q^*_N-Q^*_A$, 45. ΔEN (from electronegativity of elements), 46. B_0 (from prediction of partial least squares)

Suh and Rajan -Materials Science and Technology (2008)

Dimensionality Reduction Strategies

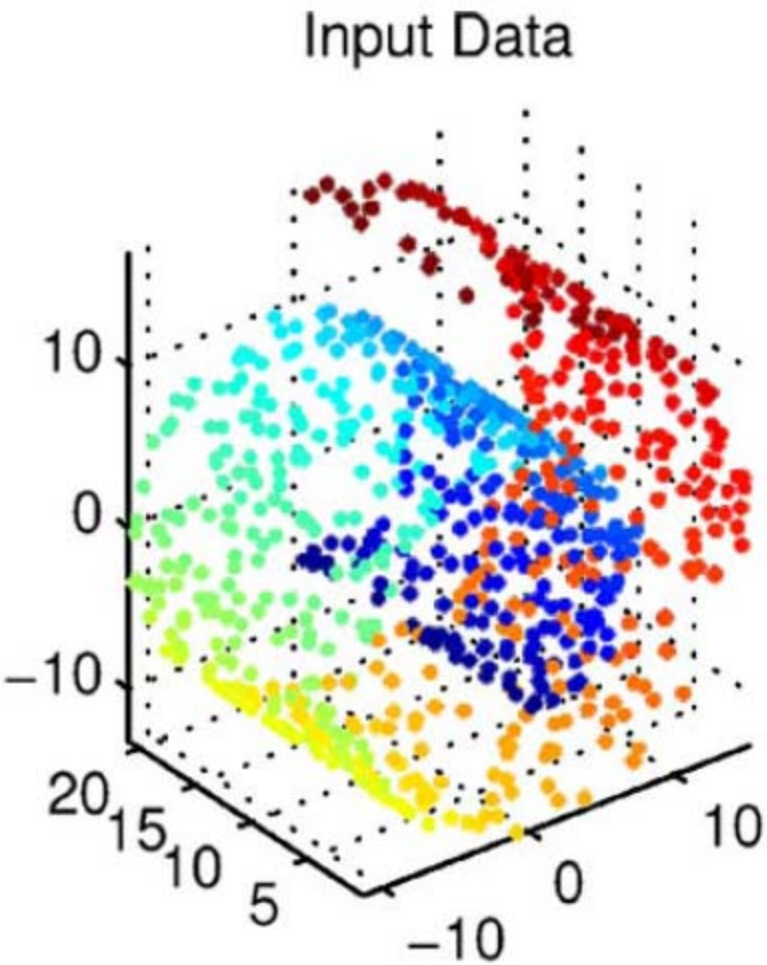


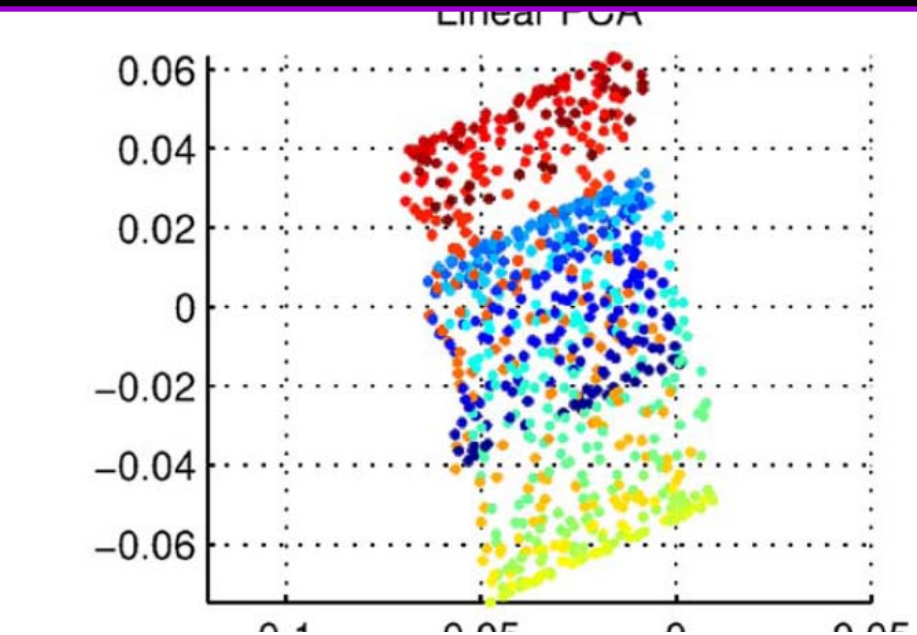
Non-linear Dimensionality Reduction

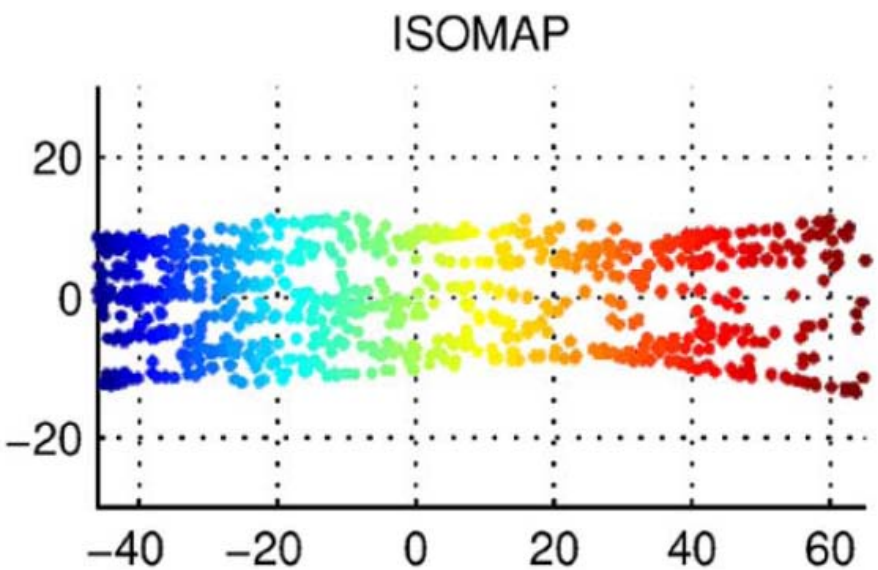


Balachandrran & Rajan (2010)

Comparing Manifolds

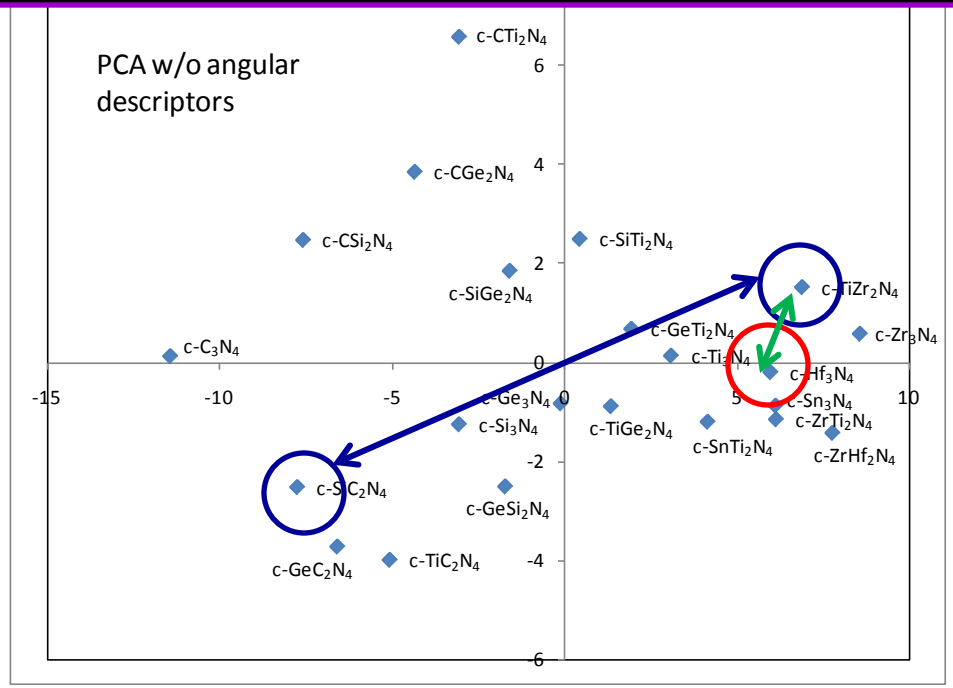


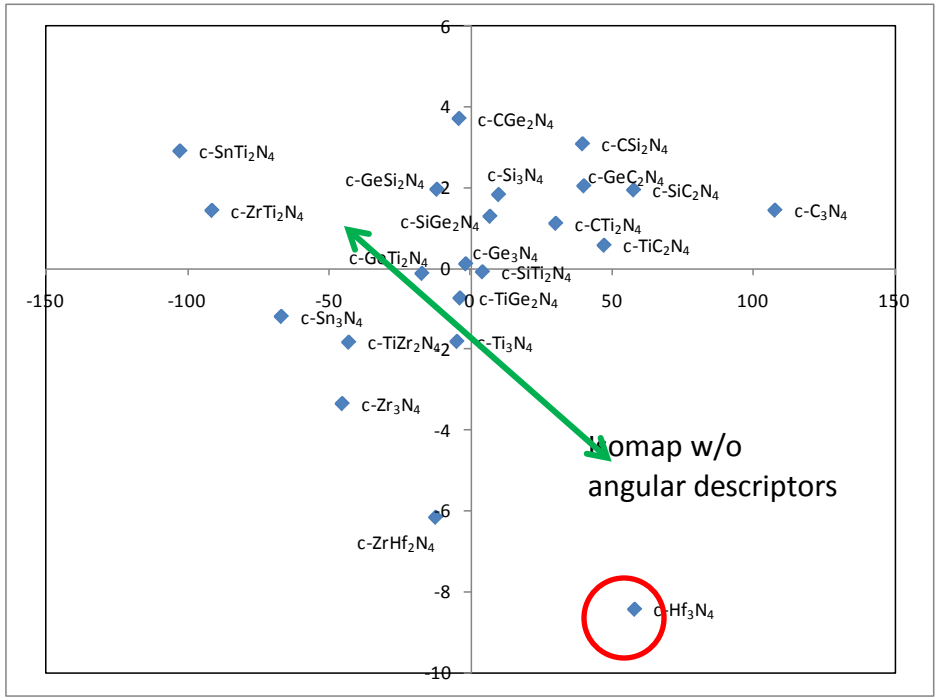




Huang et.al. IEEE-Cybernetics- 2009

Comparison of Similarity Metrics





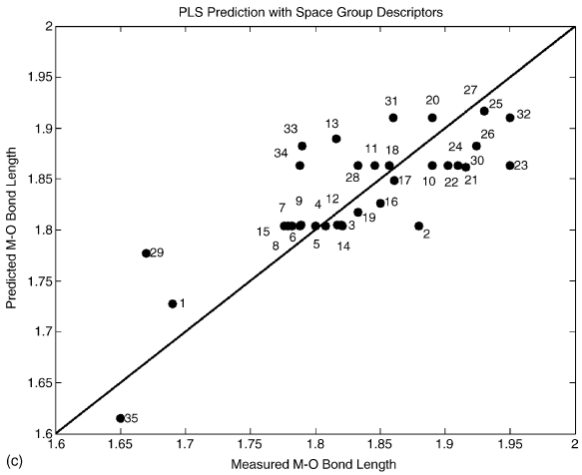
Rajan (2010)

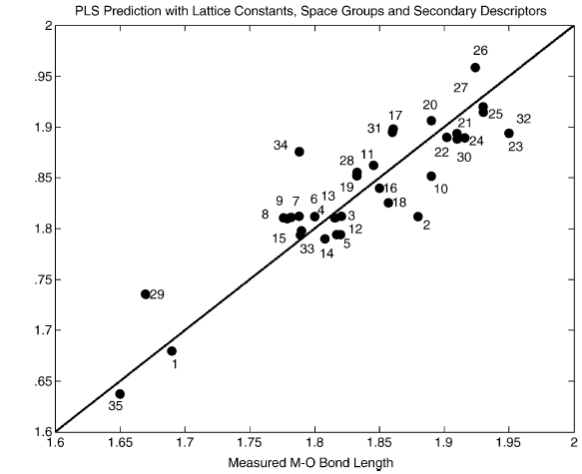
Summary and Prospects

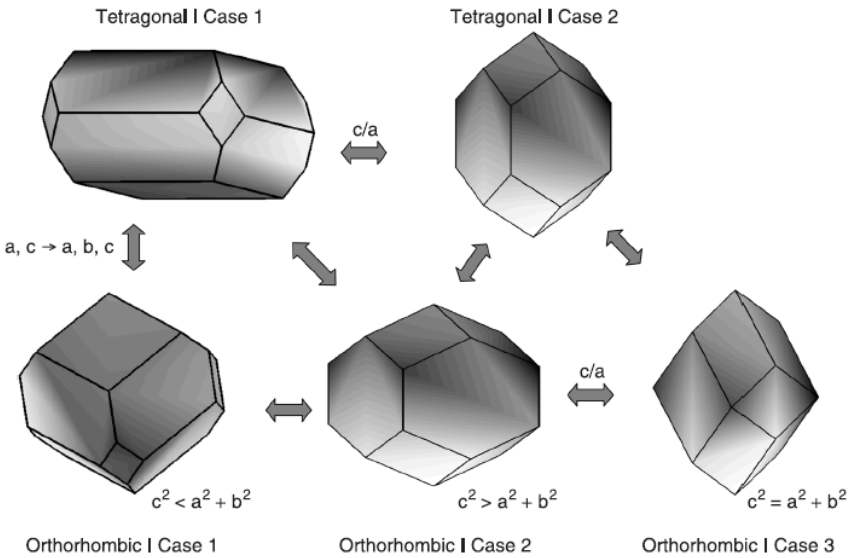
-Identification of new crystal chemistries for experiments

-Dynamic experiments for newly identified structure-bonding relationships

- Creation of data based simulations







Rajagopalan and Rajan. CRE Press -2005

Rajagopalan et.al . Applied Catalysis-2003

Acknowledgements

National Science Foundation
WHERE DISCOVERIES BEGIN

- Cyber-enabled Discovery and Innovation (CDI 2)
- Academic Research Initiative
- Algorithmic Foundations – CISE

DARPA

Center on Interfacial Engineering for Microelectromechanical Systems

Homeland Security