

Lecture 1 : Unsupervised Learning

Scott Broderick

Department of Materials Design and Innovation
University at Buffalo

September 15, 2026

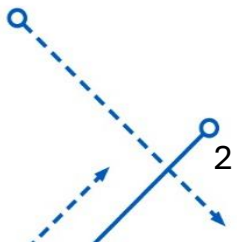


www.mdi.buffalo.edu

- *Knowledge discovery in materials science by harnessing the tools of information and data science*
- *Accelerated design and discovery of new materials, emergent properties and processing strategies that impact a broad range of technologies*

❖ > 10 years since department started

❖ 1st department built entirely on data driven design and discovery of materials & Materials Informatics





University at Buffalo

Department of Materials
Design and Innovation

School of Engineering and Applied Sciences
College of Arts and Sciences

Department of Materials Design and Innovation

mdi.buffalo.edu

Offer both Masters and
PhD Degrees.

Masters degree can be
completed in 12 months.

Courses involve
fundamental materials
science, machine learning
statistics, design of
experiments.



School of Engineering and Applied Sciences || College of Arts and Sciences

Department of Materials Design and Innovation

Give to MDI

Apply Now

About Us

Academics

Research

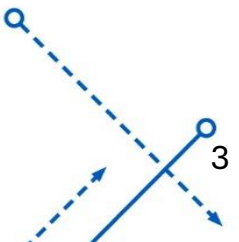
People

News and Events



Department of Materials Design and Innovation > Academics > Graduate

Graduate Education



Overview

This is the first in a 4 lecture series.

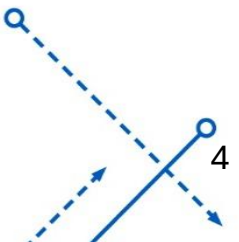
(1) Unsupervised Learning

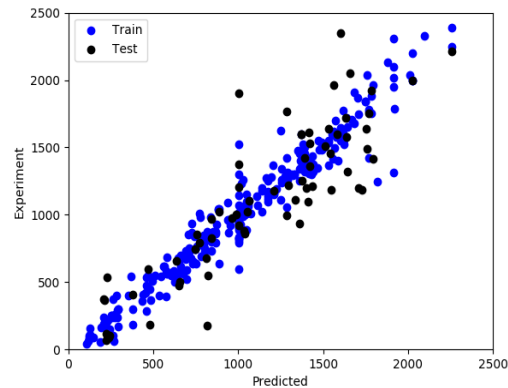
(2) Supervised Learning --- with relatively little data

(3) Machine Learning (e.g. Neural Networks) ---- with lots of data

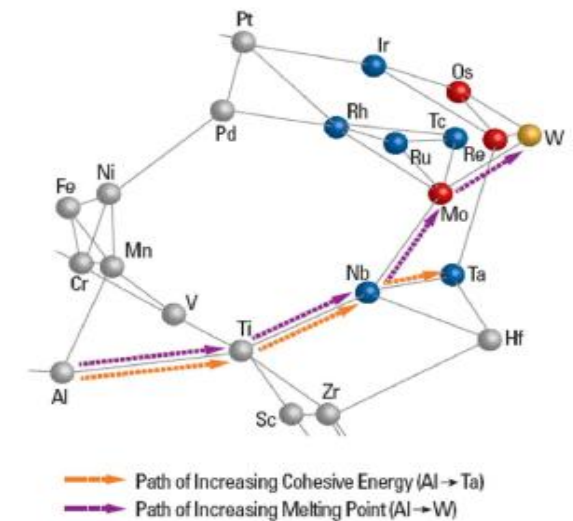
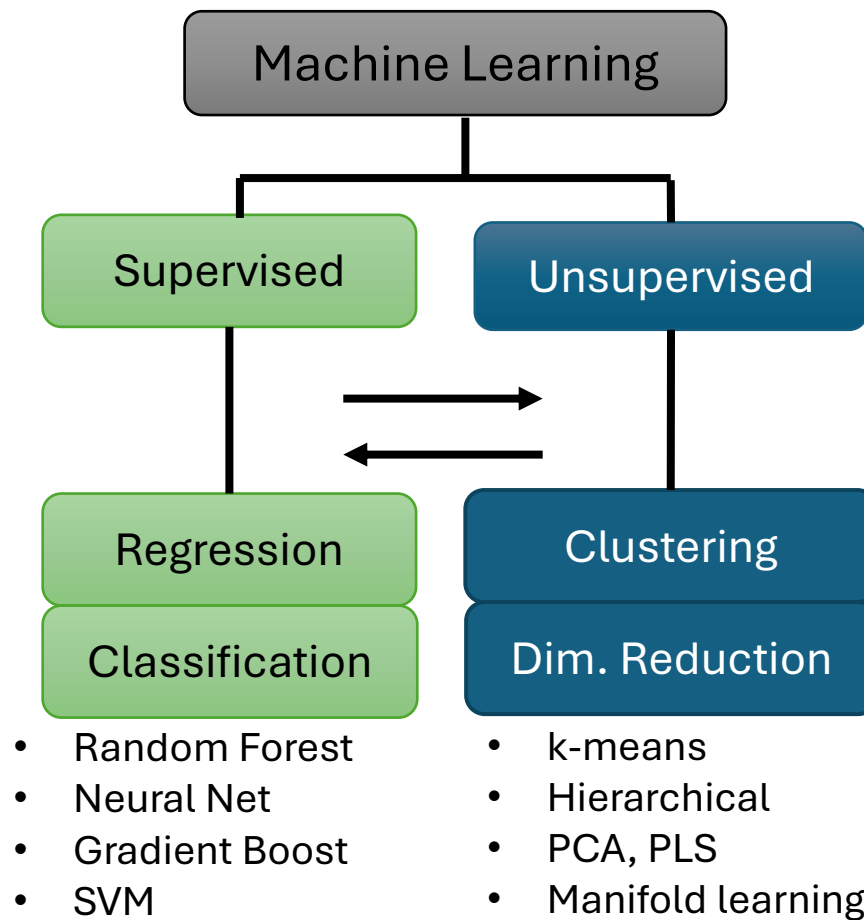
(4) Application and usage (Python)

Today's lecture will cover a couple techniques in unsupervised learning (primarily principal component analysis and IsoMap) with brief mathematical background and focus on a few examples and the interpretations. We will start with discussing data and end with a brief demonstration of analysis in R.

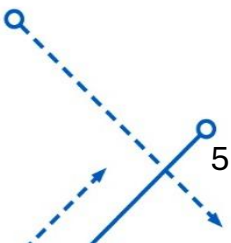




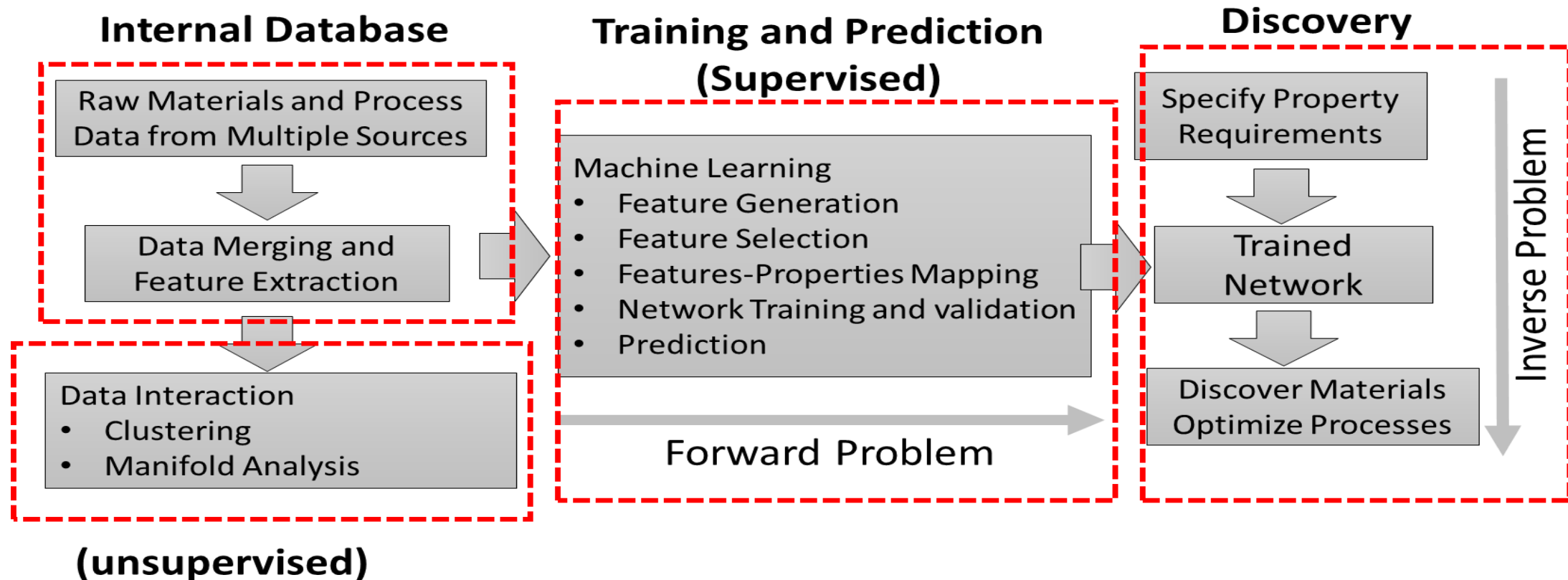
- Quantitative Prediction
- Inverse Design



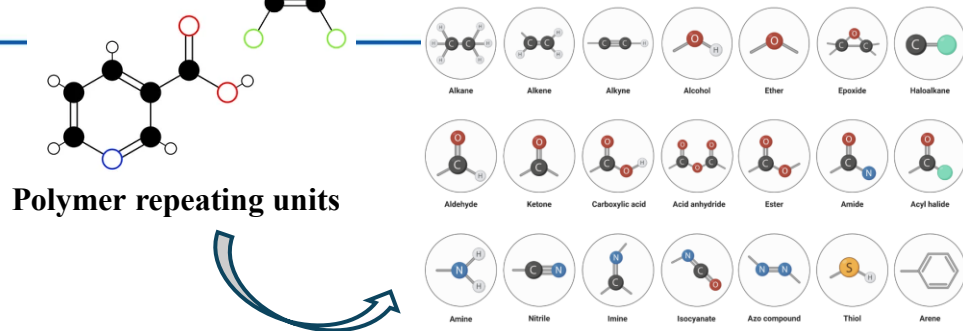
- Data Connection
- Physics
- Design Rules



Machine Learning (ML) Design Logic

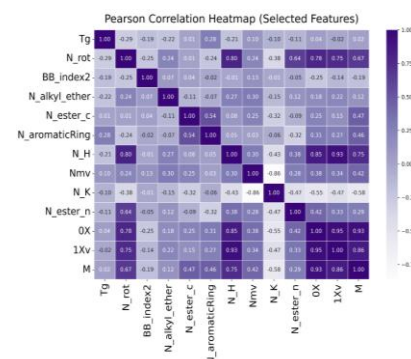


Example Study Framework



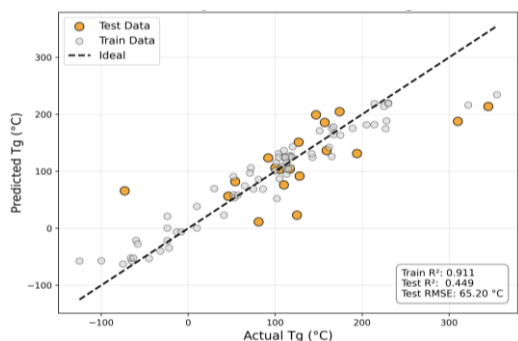
Develop **topological descriptors** for modeling and design-rule discovery

Design polymers with tailored mechanical, thermal, and chemical properties.



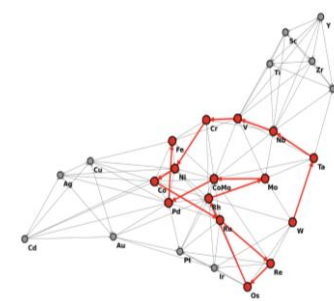
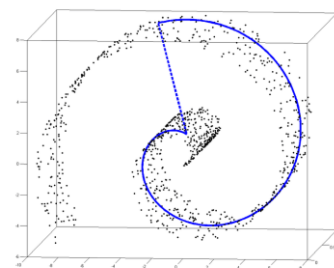
Correlation Study
(Build Structure-Property Relationships, Identify what molecular features are important by polymer topological data)

ML Models
(Predict Key Performance Metrics)



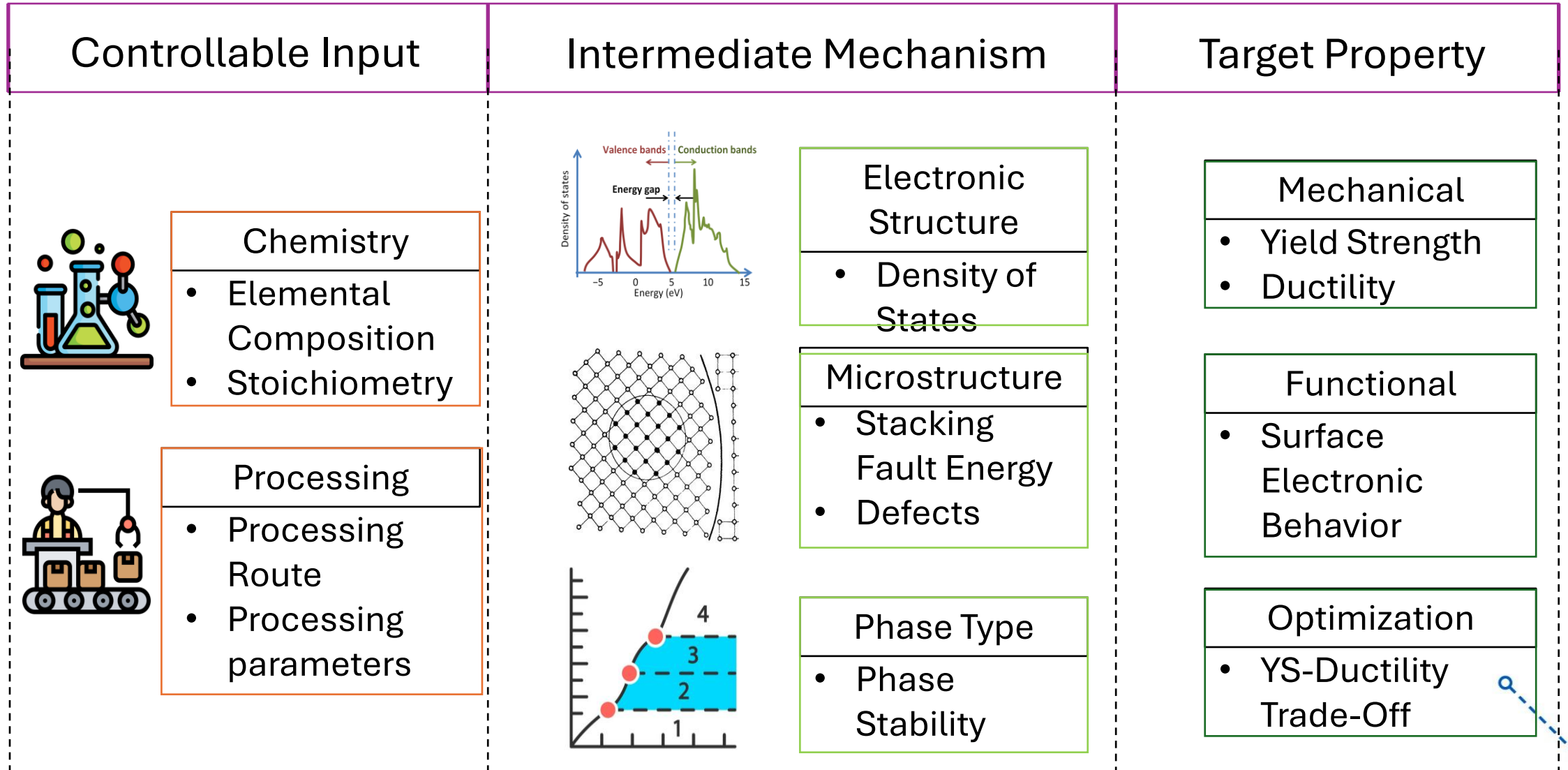
Establish **structure-property relationships** linking polymer chemistry to key performance metrics (T_g , Elongation, Corrosion Resistance, etc.).

Complement ML predictions with **network-based exploration** to identify promising candidates under **limited data**.



Manifold Learning & Random Walk Simulation
(identify hidden patterns and relationships between molecular features and polymer properties, increase prediction accuracy for limited data size)

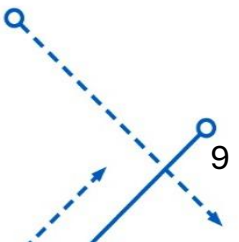
Linking Length Scales, Controls and Properties



In the different analyses we will do, there are a lot of considerations of data that we will need to consider. We will discuss those within the context of the different case studies, and understand what we do and when.

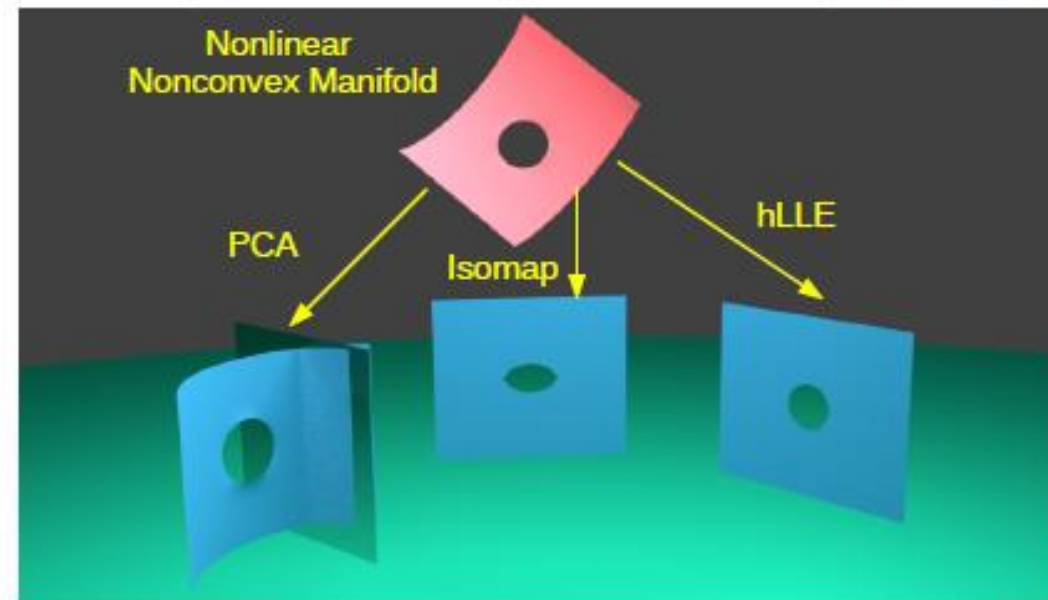
Some issues include the following :

- missing or incomplete data
- challenges associated with small data and challenges associated with large data
- outliers
- uncertainty
- normalization
- multiple overlapping measurements
- inter-correlations in the data
- skewing or bias in the data



Manifolds / Data Projections

	PCA	Isomap	hLLE
Assumptions	Linear	Convex	Locally Linear
Property Preserved	Euclidean Distance	Geodesic Distance	Local Linearity
Intuitive Description	Fits linear plane	Rolls it out	•Subdivides M •Reorients pieces •Stitches the pieces together

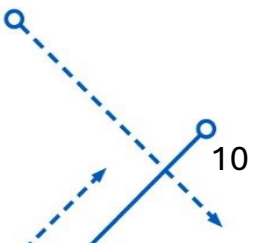


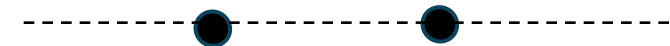
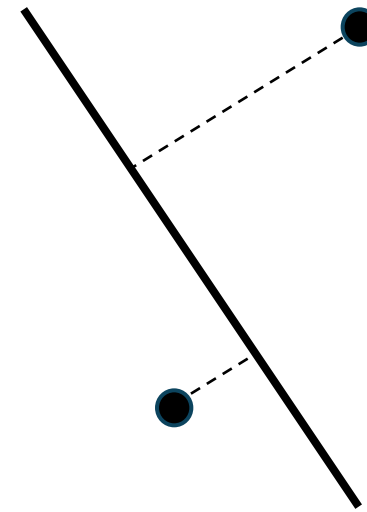
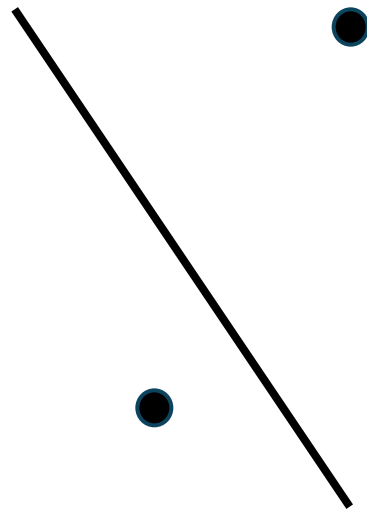
PCA fits a linear plane

Isomap extracts distorted convex corners

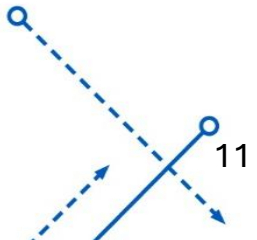
Hessian LLE

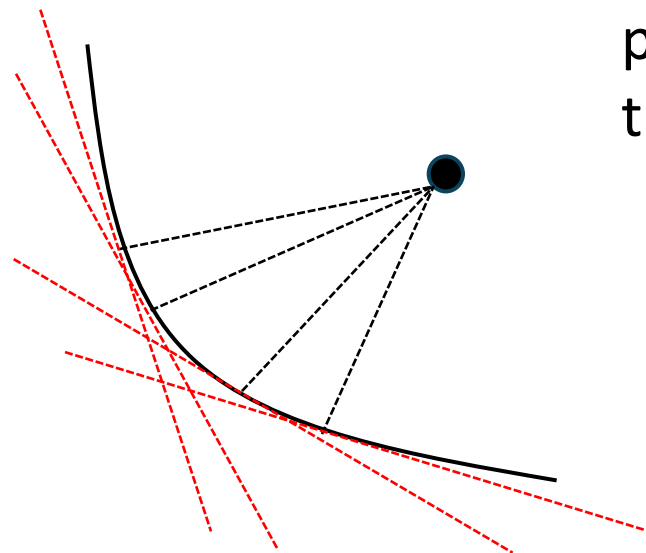
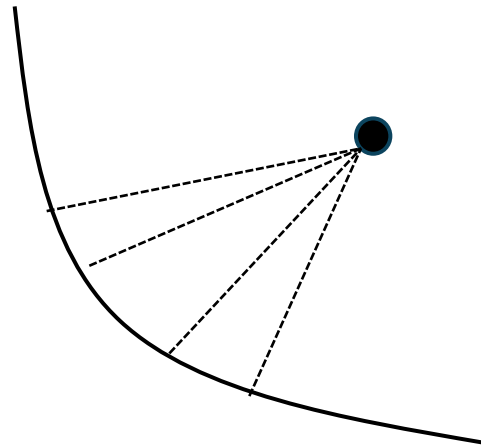
How do we define the projection ? ... how do we define the manifold ?





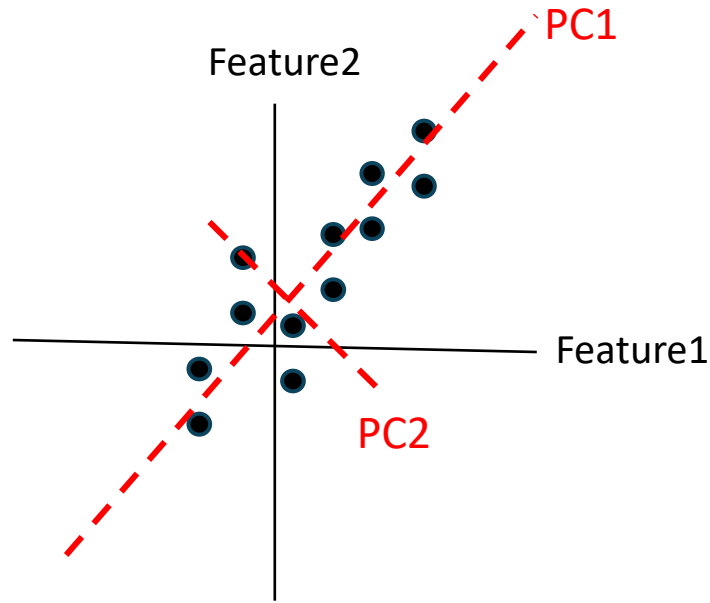
- When projecting onto a line, the choice of projection is clear Normal/Perpendicular to the line
- Logic continues to a plane Even a n -dimensional plane with large n





It becomes less clear when we have a curved line Lots of normal points Most common options are (i) the selection of the shortest normal to the line or (ii) choosing the normal line and the tangent at that to be perpendicular (ie. dot product of the two lines is equal to 0).

Principal Component Analysis (PCA)



$$X = T * P^T + E$$

Loadings (P) are the eigenvectors of the covariance matrix of the pre-processed data, ordered by eigenvalue.

We will use two matrices coming out of this : (i) the scores (T) which describe relationships in systems (original data converted into new axis system) and (ii) loadings which describe relationships in features / properties / etc. (defining the new axes system)

- In the original data, every feature has initially equal importance (ie. need n-dimensional space to describe n features)
- However, with PCA we define new dimensions (principal components : PCs) which are a linear combination of the original features, defined so as to maximize variance captured in each PC.

PCA is an unsupervised learning approach, but it can also serve multiple purposes for us in a supervised learning analysis :

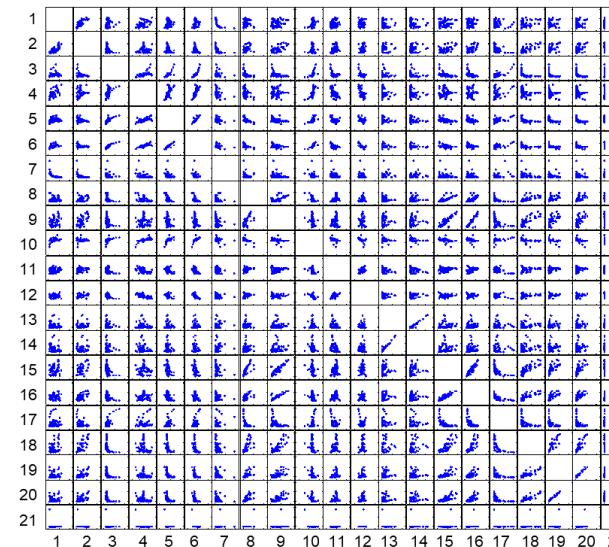
- Another representation of the data and the correlations in the data
- New parameterization of data
- Reduction of data dimensionality
- Can be used to minimize and test the descriptors kept for modeling

Pair-Wise Correlations

✓ Elemental properties used for bivariate plot

PERIODIC TABLE OF ELEMENTS

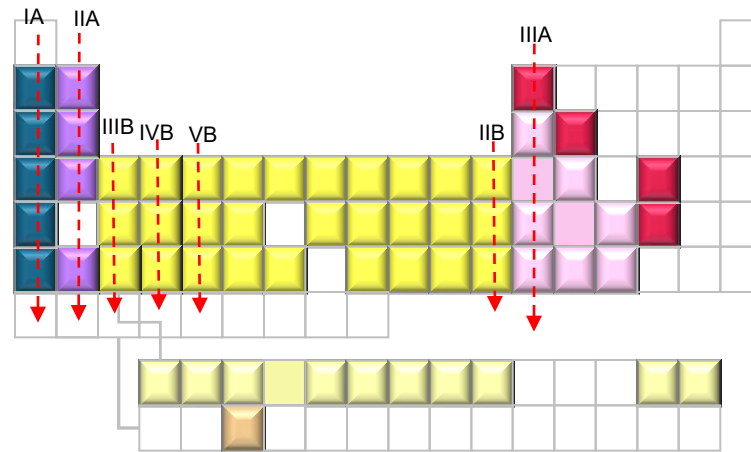
1 H Hydrogen Nonmetal																	2 He Helium Noble Gas						
3 Li Lithium Alkali Earth Metal	4 Be Beryllium Alkali Earth Metal																	5 B Boron Metalloid	6 C Carbon Nonmetal	7 N Nitrogen Nonmetal	8 O Oxygen Nonmetal	9 F Fluorine Halogens	10 Ne Neon Noble Gas
11 Na Sodium Alkali Earth Metal	12 Mg Magnesium Alkali Earth Metal																	13 Al Aluminum Post Transition Metal	14 Si Silicon Metalloid	15 P Phosphorus Nonmetal	16 S Sulfur Nonmetal	17 Cl Chlorine Halogens	18 Ar Argon Noble Gas
19 K Potassium Alkali Earth Metal	20 Ca Calcium Alkali Earth Metal	21 Sc Scandium Transition Metal	22 Ti Titanium Transition Metal	23 V Vanadium Transition Metal	24 Cr Chromium Transition Metal	25 Mn Manganese Transition Metal	26 Fe Iron Transition Metal	27 Co Cobalt Transition Metal	28 Ni Nickel Transition Metal	29 Cu Copper Transition Metal	30 Zn Zinc Transition Metal	31 Ga Gallium Post Transition Metal	32 Ge Germanium Metalloid	33 As Arsenic Metalloid	34 Se Selenium Nonmetal	35 Br Bromine Halogens	36 Kr Krypton Noble Gas						
37 Rb Rubidium Alkali Earth Metal	38 Sr Strontium Alkali Earth Metal	39 Y Yttrium Transition Metal	40 Zr Zirconium Transition Metal	41 Nb Niobium Transition Metal	42 Mo Molybdenum Transition Metal	43 Tc Technetium Transition Metal	44 Ru Ruthenium Transition Metal	45 Rh Rhodium Transition Metal	46 Pd Palladium Transition Metal	47 Ag Silver Transition Metal	48 Cd Cadmium Transition Metal	49 In Indium Post Transition Metal	50 Sn Tin Post Transition Metal	51 Sb Antimony Metalloid	52 Te Tellurium Metalloid	53 I Iodine Halogens	54 Xe Xenon Noble Gas						
55 Cs Cesium Alkali Earth Metal	56 Ba Barium Alkali Earth Metal	57 La Lanthanum Lanthanide	58 Ce Cerium Lanthanide	59 Pr Praseodymium Lanthanide	60 Nd Neodymium Lanthanide	61 Pm Promethium Lanthanide	62 Sm Samarium Lanthanide	63 Eu Europium Lanthanide	64 Gd Gadolinium Lanthanide	65 Tb Terbium Lanthanide	66 Dy Dysprosium Lanthanide	67 Ho Holmium Lanthanide	68 Er Erbium Lanthanide	69 Tm Thulium Lanthanide	70 Yb Ytterbium Lanthanide	71 Lu Lutetium Lanthanide							
87 Fr Francium Alkali Earth Metal	88 Ra Radium Alkali Earth Metal	89 Ac Actinium Actinide	90 Th Thorium Actinide	91 Pa Protactinium Actinide	92 U Uranium Actinide	93 Np Neptunium Actinide	94 Pu Plutonium Actinide	95 Am Americium Actinide	96 Cm Curium Actinide	97 Bk Berkelium Actinide	98 Cf Californium Actinide	99 Es Einsteinium Actinide	100 Fm Fermium Actinide	101 Md Mendelevium Actinide	102 No Nobelium Actinide	103 Lr Lawrencium Actinide							



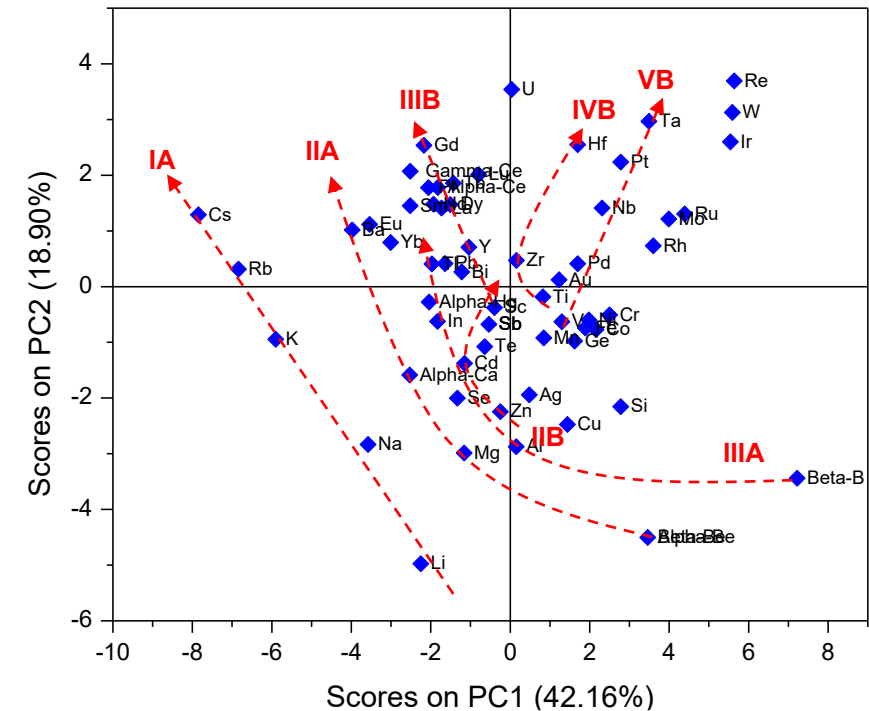
1. Atomic Weight
2. Density @ 293 K
3. Molar Volume
4. Pseudopotential core radii sum
5. Atomic Radius
6. Covalent Radius
7. Specific Heat
8. Heat of Fusion
9. Heat of Vaporization
10. Heat Capacity
11. the Martynov-Batsanov electronegativity
12. First Ionization Potential
13. Electrical Conductivity
14. Thermal Conductivity at 300K
15. Melting Point
16. Boiling Point
17. Coefficient of Thermal Expansion
18. Bulk Modulus
19. Shear Modulus
20. Modulus of Elasticity
21. Electrical Resistivity

Scores

Structure Sequence of the Periodic Table by PCA - Patterns for Columns (or Groups)



✓ PCA Score Plot (Space for elements)



Diagonal patterns (red arrows) on 2D space represent each column of the periodic table.

This provides a more accurate representation of the periodic table and the relationship of elements..... While maintaining the basic organization of the periodic table

Loadings

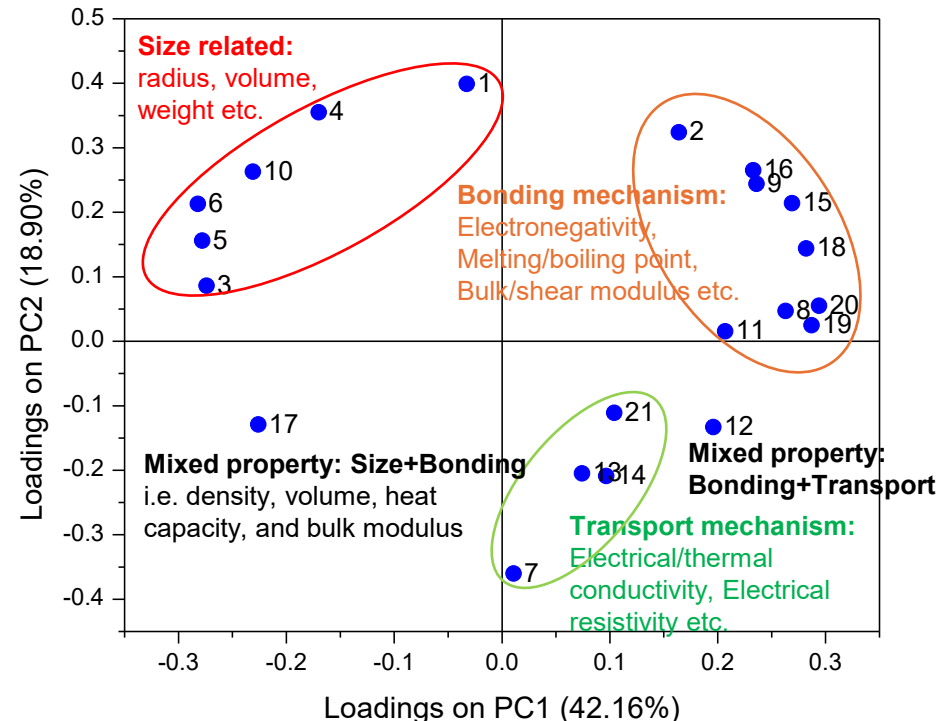
Prediction of Structure-Property Relationships by PCA

✓ Elemental properties used in the 2nd PCA test

1. Atomic Weight
2. Density @ 293 K
3. Molar Volume
4. Pseudopotential core radii sum
5. Atomic Radius
6. Covalent Radius
7. Specific Heat
8. Heat of Fusion
9. Heat of Vaporization
10. Heat Capacity
11. the Martynov-Batsanov electronegativity
12. First Ionization Potential
13. Electrical Conductivity
14. Thermal Conductivity at 300K
15. Melting Point
16. Boiling Point
17. Coefficient of Thermal Expansion
18. Bulk Modulus
19. Shear Modulus
20. Modulus of Elasticity
21. Electrical Resistivity

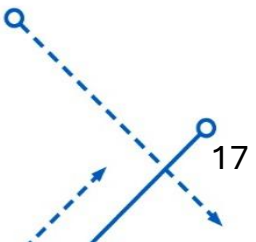
✓ PCA Loading Plot (Space for properties)

c.f. each number on this figure corresponds to elemental properties (left).



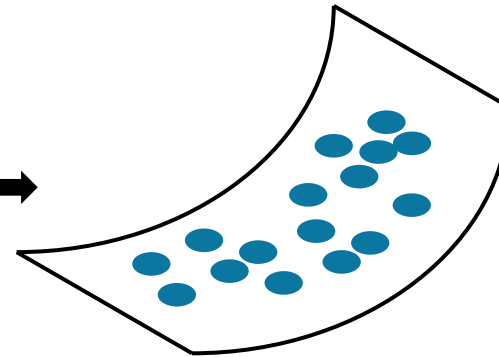
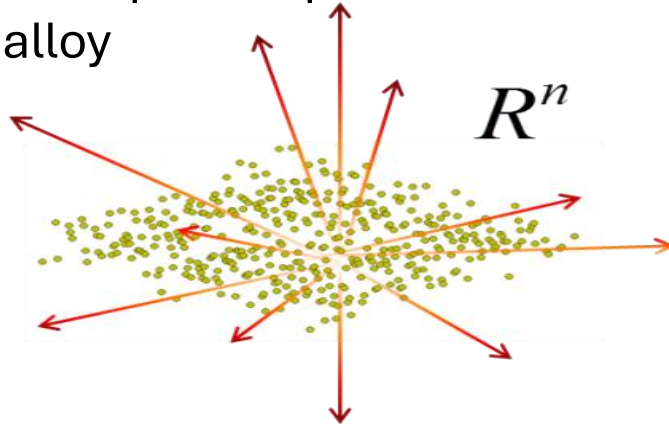
What do we learn?

PCA loading plot represents overall correlations between properties. PCA loading plot produces clusters based on similar characteristics of properties (i.e. size, bonding mechanism, and transport mechanisms). Since #17 (coefficient of thermal expansion) and bulk modulus (#18) or #17 and melting point (#15) sit diagonally, they have negative relationships (i.e. dissociation vs. association in bonding).

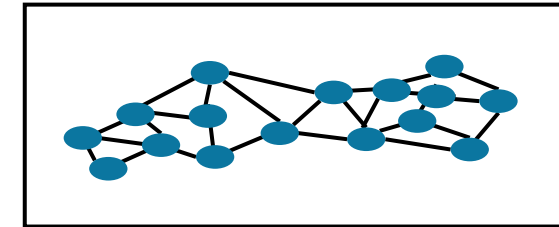


Converting Data to a Graph

(1) n features / descriptors
Each point represents an alloy



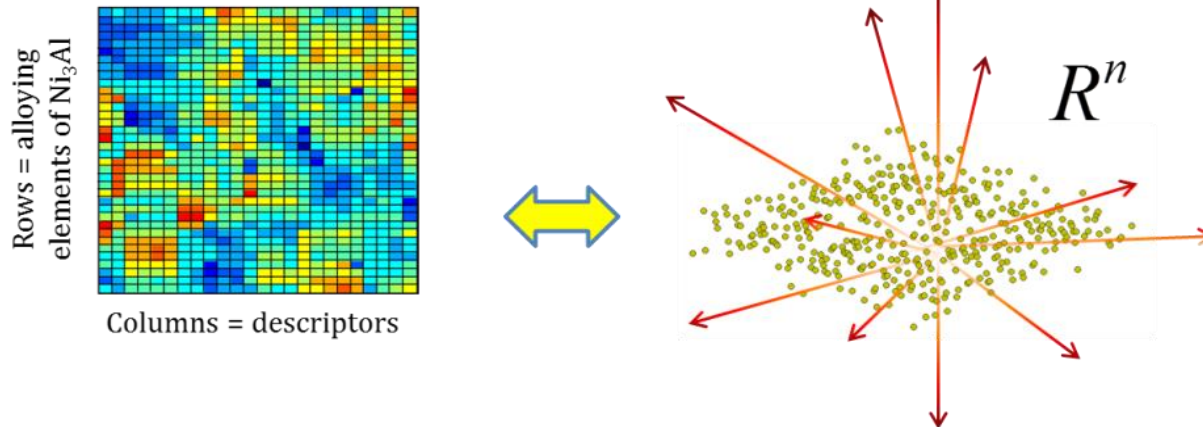
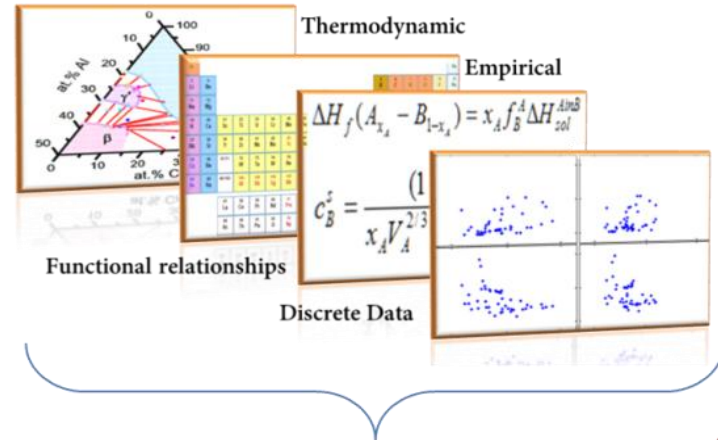
(3) 'flatten' the manifold and define k nearest connections by geodesic distance



(2) Identify a reduced dimension non-linear manifold which maximizes information captured and project data onto the manifold

- Points which are connected are more similar than points that are not connected. The number of connections required to connect points ("pathways") is inverse to similarity.
- Within connected points, the shorter the connection then the higher the similarity.
- The network adapts with new data. Therefore, can be used to assess sensitivity to data.

IsoMap



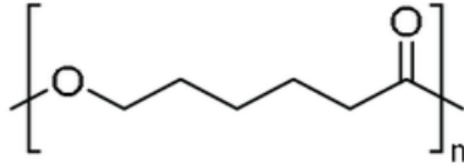
Each descriptor is represented by a dimension in R^n .

Each alloying element is represented by a data point in this space.

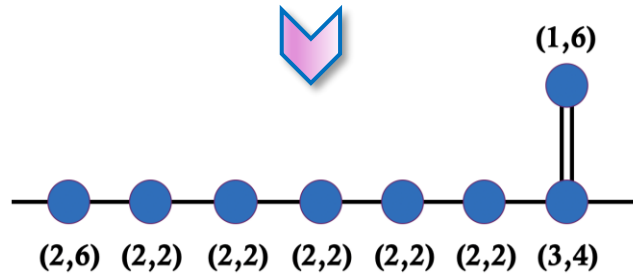
➤ To develop a dimensionally reduced mapping of the data, we employ a non-linear manifold approach.

1. Computation of the Euclidean distance matrix $[E]$ between each pair of points (x_i, x_j) in R^n .
2. Retention of only those entries of $[E]$ which connect every point to its k nearest neighbors to obtain the reduced matrix $[E_k]$.
3. Calculation of the geodesic distances matrix $[G]$ by approximating with the graph distances between each pair of points (x_i, x_j) in R^n from $[E_k]$.
4. Computation of the dissimilarity matrix $[A]$.
5. Solving the eigenvalue problem for $[A]$ to obtain the eigenpairs (k, v) . The product of the eigenpairs provide the low-dimensional embedding of the nonlinear manifold.

Polymer Topological Data Feature Set Generation



Structure of the repeat unit of Poly (Caprolactone)



Hydrogen suppressed graph of the monomer

Basic group	δ	δ^v	Basic group	δ	δ^v
-CH ₃	1	1	-NH ₂	1	3
-CH ₂ -	2	2	-N-	2	5
=CH ₂	1	2	-N<	3	5
-CH<	3	3	-OH	1	5
>C<	4	4	=O	1	6
=CH-	2	3	-O-	2	6
=C<	3	4	-F	1	7
-NH-	2	4	-Cl	1	7/9

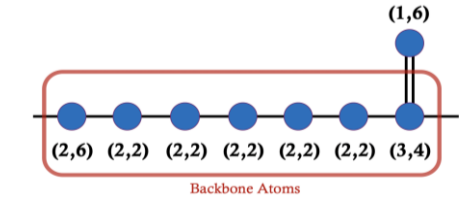
Assign (δ , δ_v) values to each atom in the repeat unit

BB_index1 $\rightarrow$ backbone flexibility (steric hindrance)

BB_index2 $\rightarrow$ local variation in backbone environment

$$\text{BB_index1} = \frac{\sum_{\text{BBatoms}} (\frac{\delta}{\delta^v})}{\text{NBB}}$$

$$\text{BB_index2} = 3 - \frac{\sum_{\text{BBatoms}} (\delta)}{\text{NBB}}$$



NBB – the number of backbone atoms

Backbone index1 and Backbone index2

$^0\chi$ $\rightarrow$ atom-level connectivity

$^1\chi$ $\rightarrow$ bond-level connectivity

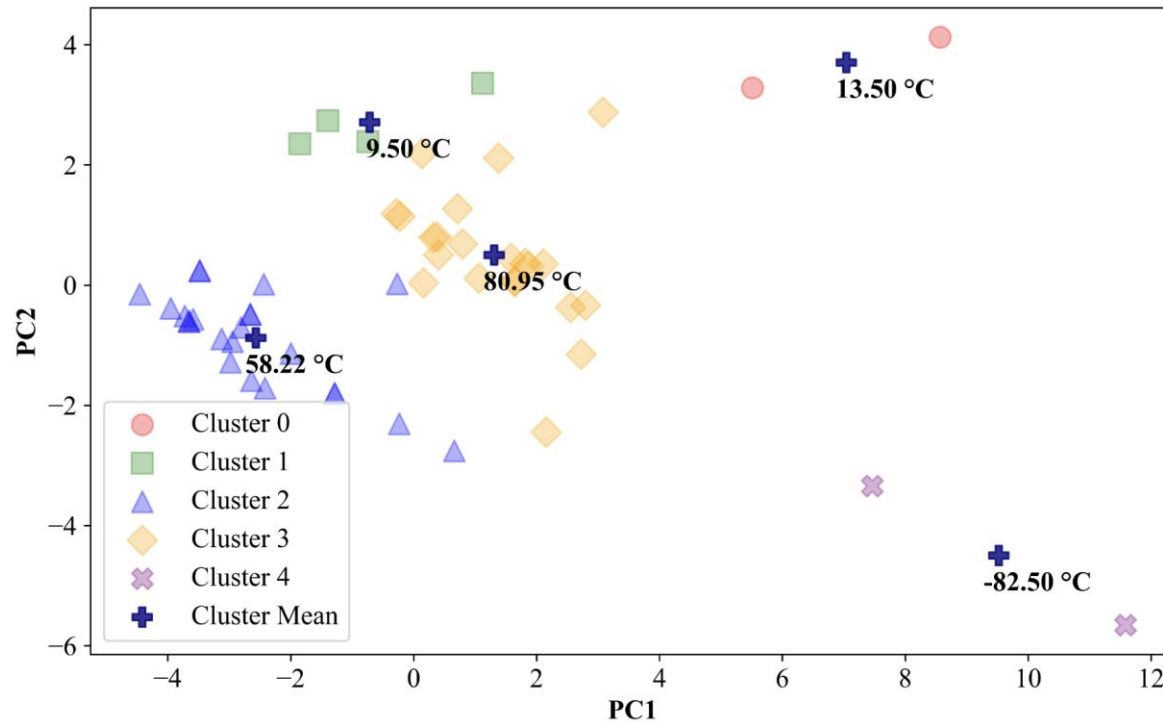
$^0\chi_v$ $\rightarrow$ atomic electronic structure

$^1\chi_v$ $\rightarrow$ bond electronic structure

$$^0\chi^v = \sum_{i \in N} \frac{1}{\sqrt{\delta_i^v}} \quad ^0\chi = \sum_{i \in N} \frac{1}{\sqrt{\delta_i}} \quad ^1\chi = \sum_{(i,i') \in E} \frac{1}{\sqrt{\delta_i \delta_{i'}}} \quad ^1\chi^v = \sum_{(i,i') \in E} \frac{1}{\sqrt{\delta_i^v \delta_{i'}^v}}$$

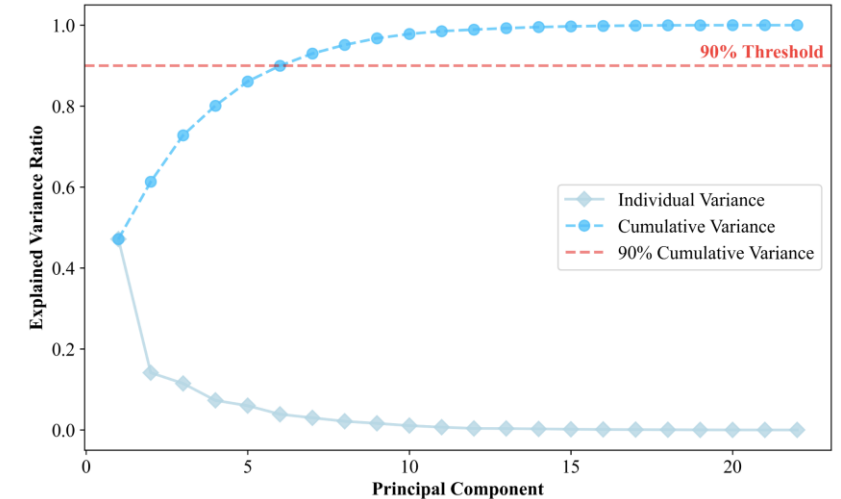
Zeroth-order and first-order (atomic) connectivity indices

Clusters and Mean Tg Values

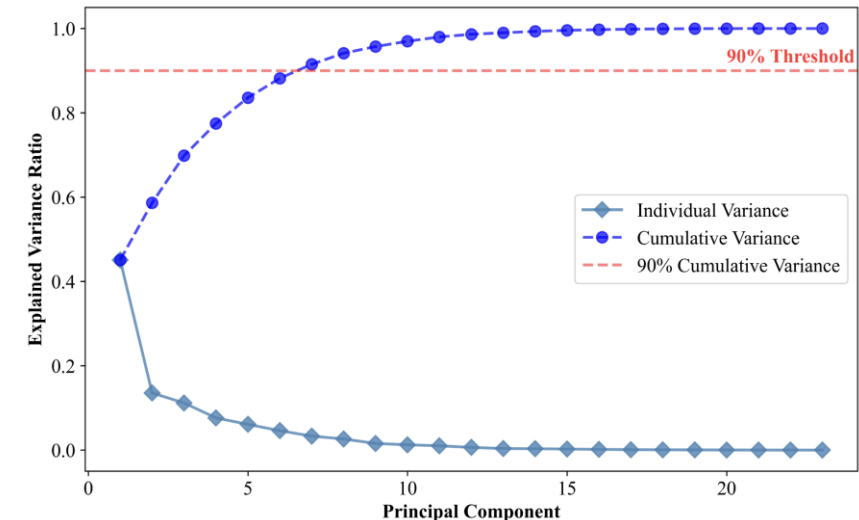


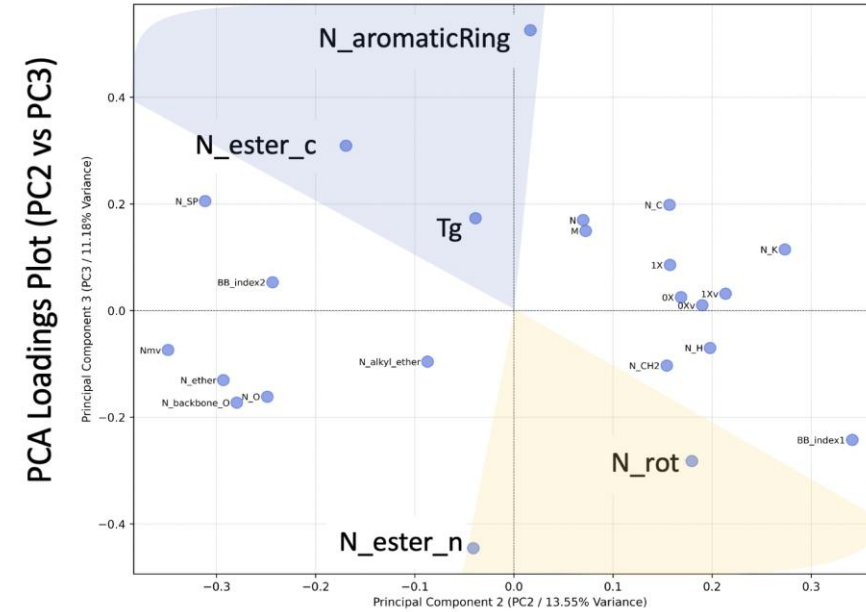
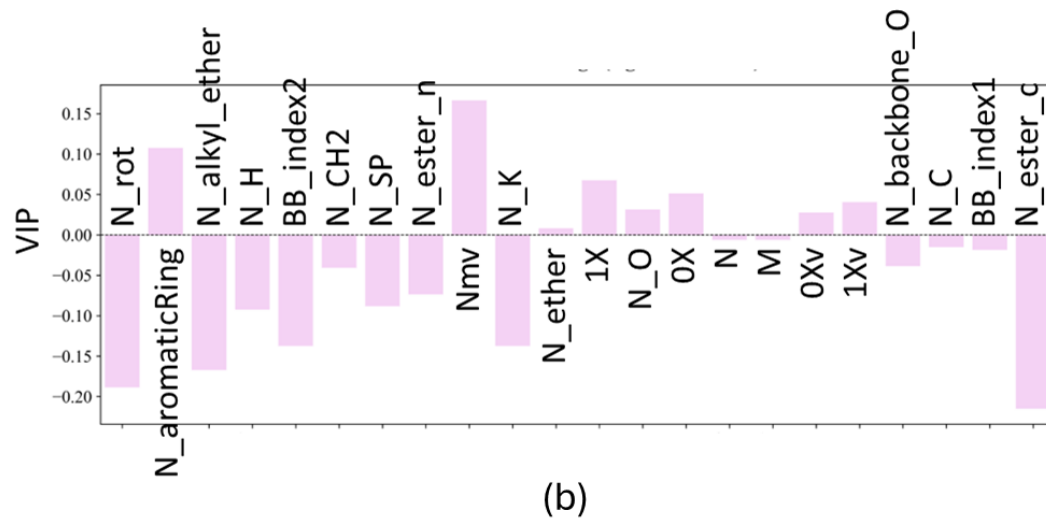
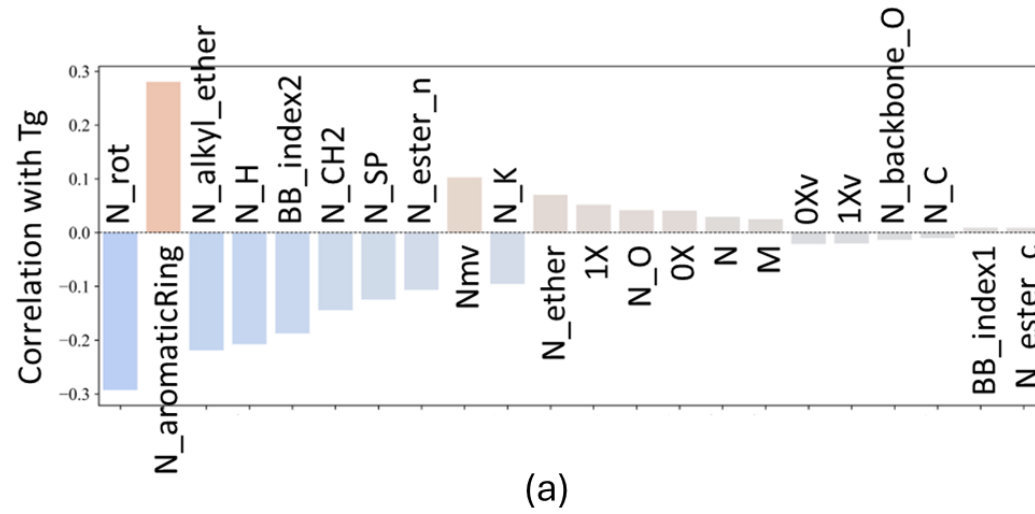
- The input into the PCA (above) was only topological descriptors. We see the high Tg clustered together without Tg as an input.
- Inclusion of Tg does not introduce significantly different variance in data, suggesting redundancy of information.
- From this, suggests that the feature set indeed captures Tg while providing an interpretable input.

Explained Variance by Principal Component (without Tg)



Explained Variance by Principal Component (with Tg)





$$\text{VIP} = \frac{\text{PC}_x^T \times \text{PC}_x^i}{\sum \text{PC}_x^T \times \text{PC}_x^i}$$

Assessing Relationship Between Features and Tg

To get scientific level information on relationship between features and Tg, four different approaches were applied to define correlation

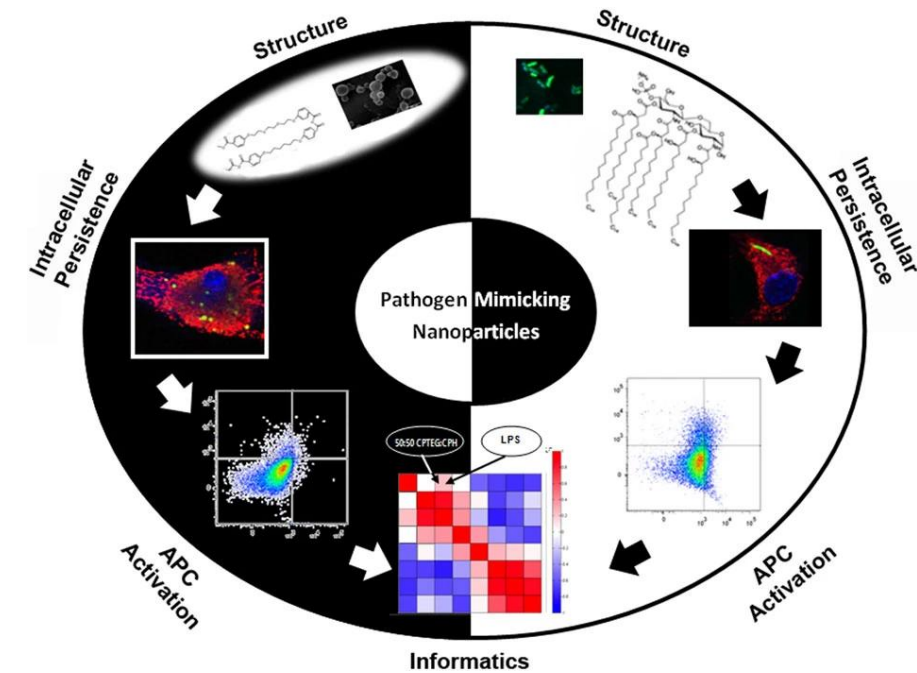
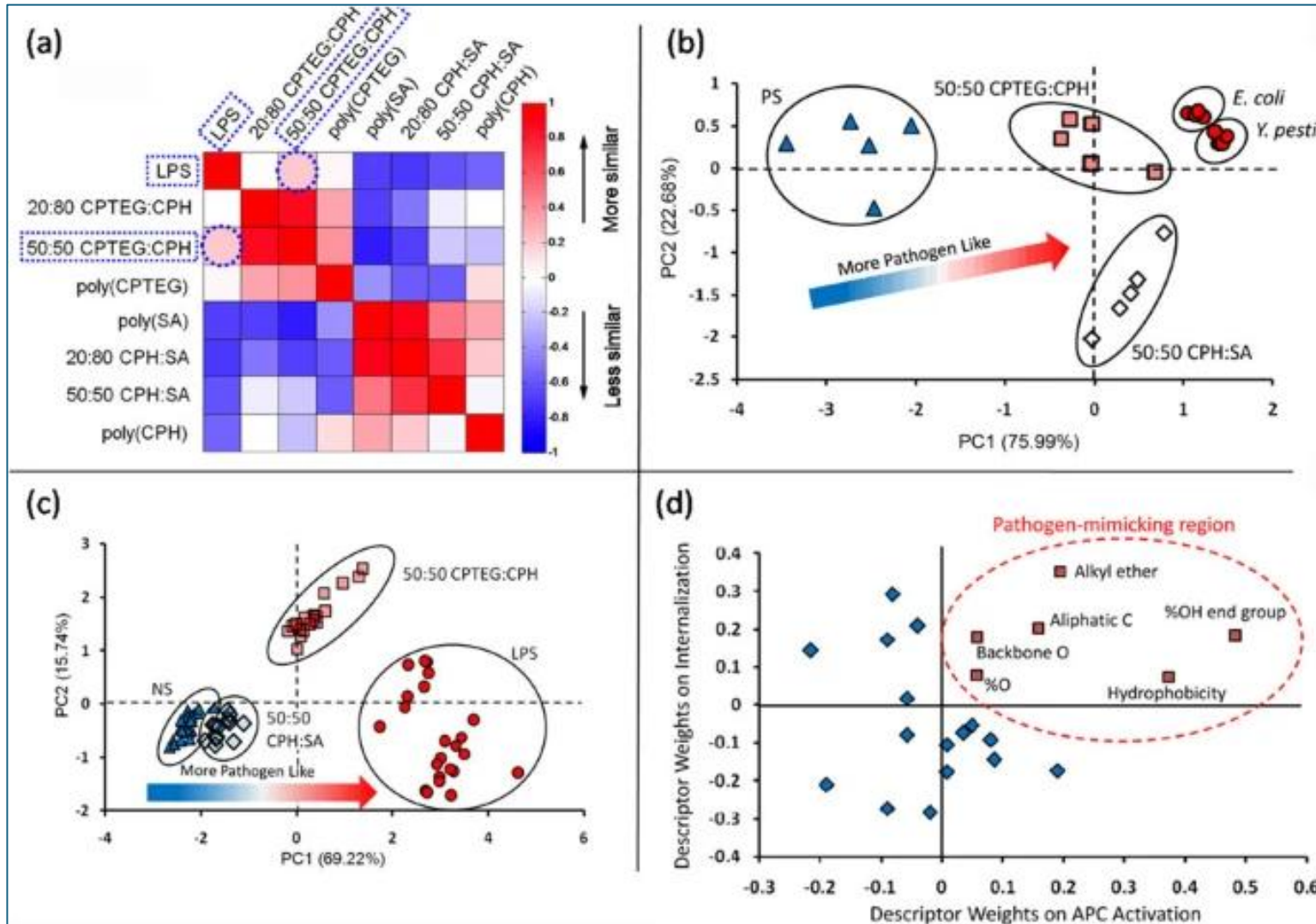
- Pair-wise correlation
- Variable importance projection (within PCA)
- 2D PCA loadings relationships
- Random Forest feature importance

The reason for applying four different methods was to ensure consistency and not due to bias in the method or some artifact in the data.

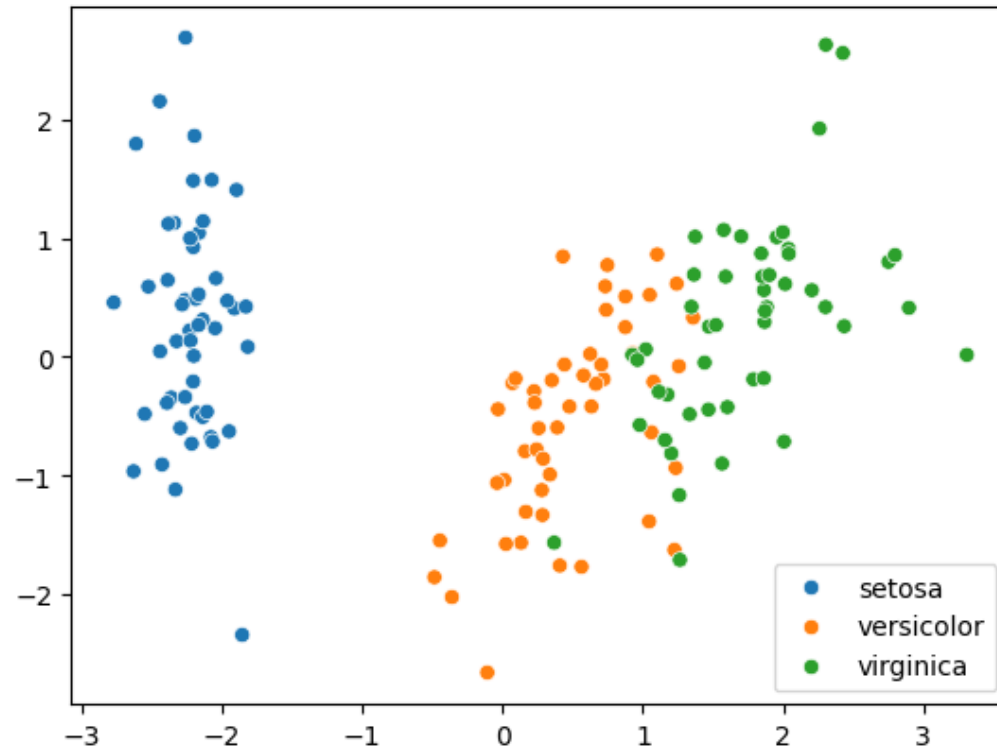
Feature	Times Selected	Correlation
N_rot	4	-
BB_index2	3	-
N_alkyl_ether	2	-
N_ether_c	2	+ and -
N_aromaticRing	2	+
N_H	2	-
Nmv	1	+
N_K	1	-
N_ester_n	1	-
0X	1	
1Xv	1	
M	1	

- N_rot: total number of rotational degrees of freedom parameter (N_rot = the backbone rotational degrees plus the side group rotational degrees)
- BB_index2: backbone index2 is a steric hindrance parameter that differentiates between backbone atoms with the same ($\delta/\delta V$) values but different δ values, reflecting variations in the number of non-hydrogen neighbors around each backbone atom.
- N refers to the number of that feature
- M: mole weight of one repeat unit

Integrating Different Data Types

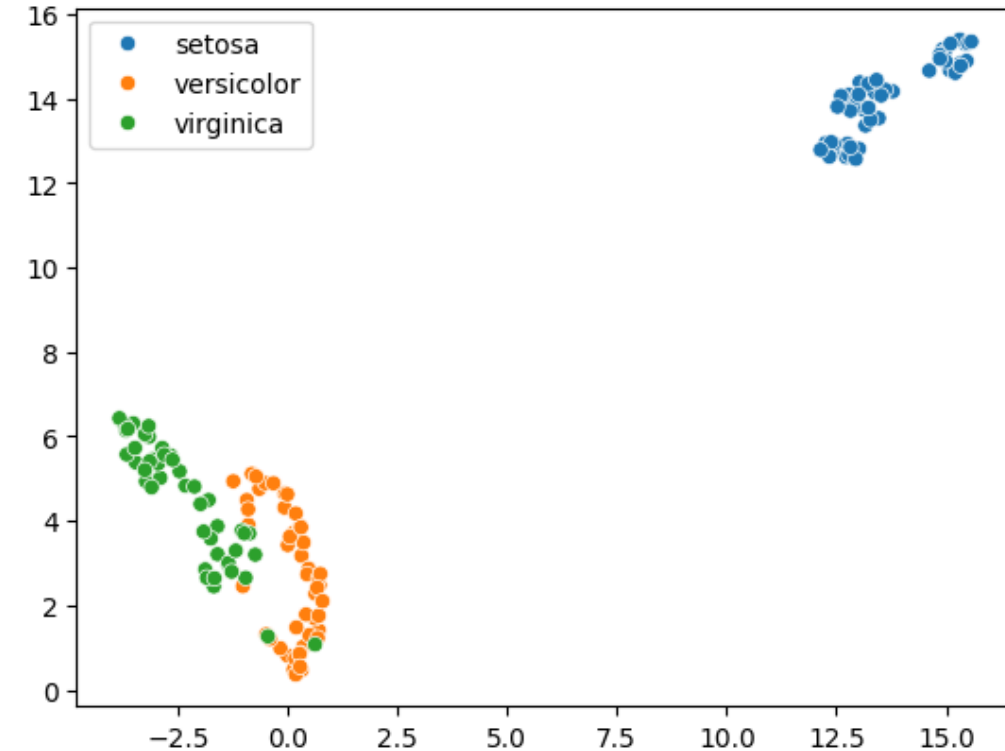


PCA on Iris Dataset

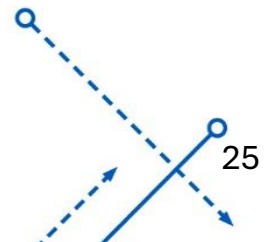


- Global representation
- Interpretable
- Assumes linearity
- Does not capture clustering as well

UMAP on Iris Dataset

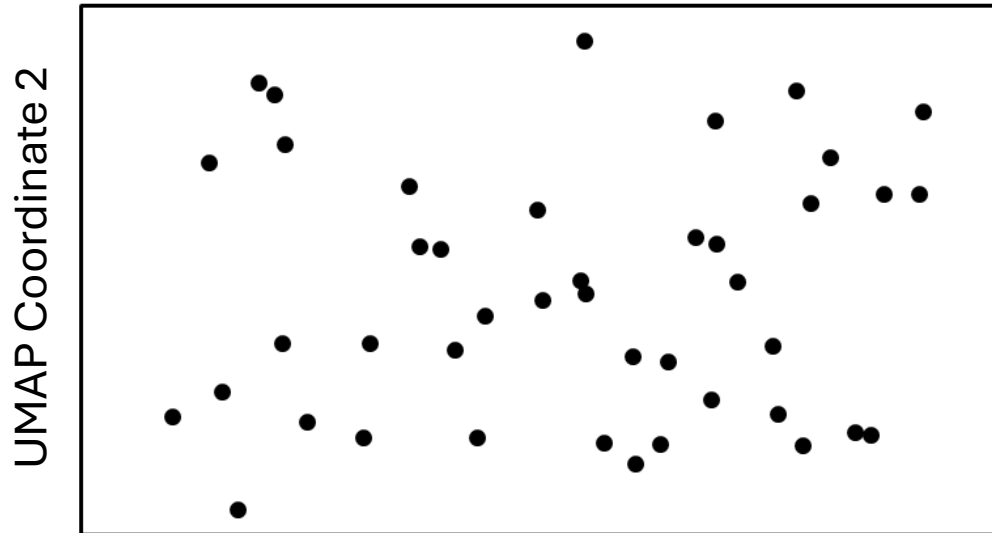


- Captures clustering
- Not as Interpretable
- Controllable parameters -- bias



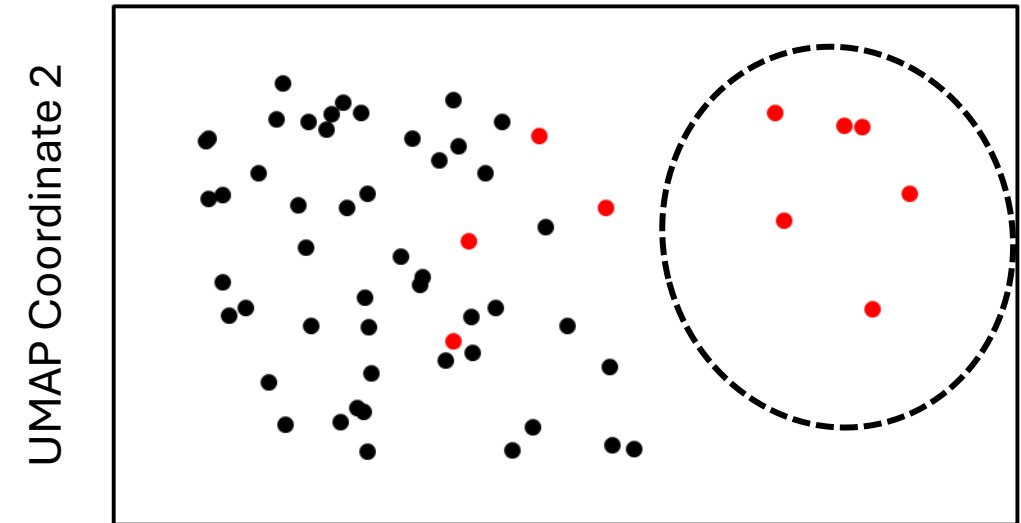
Mapping of Data Distributions

HEA Legacy Systems



UMAP Coordinate 1

With Addition of New Chemistries (Red)



UMAP Coordinate 1

- UMAP represents the distribution of the data ... that is, are the systems representing new information or providing redundant information. From mapping of original data, we see decent distribution, which provides a more robust modeling input.
- However, when adding new chemistries proposed, we find that it represents a significantly different physical relationship than others This approach leads to identification of a new space to explore which may behave different.

Sources of Bias

Bias

Pair-wise Correlations :

None

Principal Component Analysis :

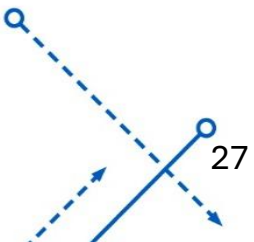
Selection of Features

Variable Importance Projection :

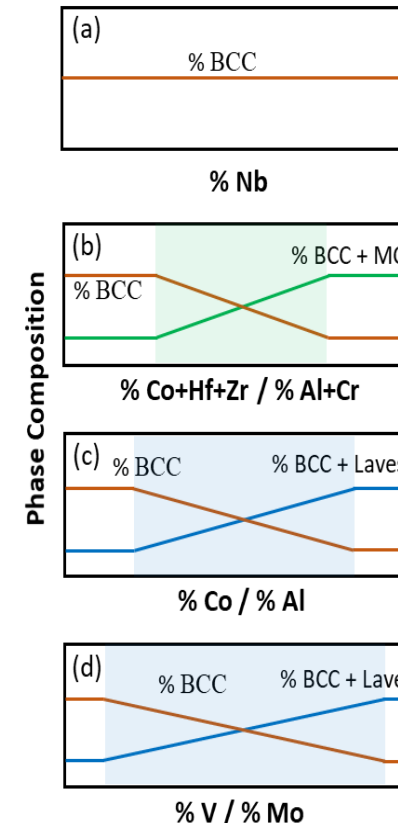
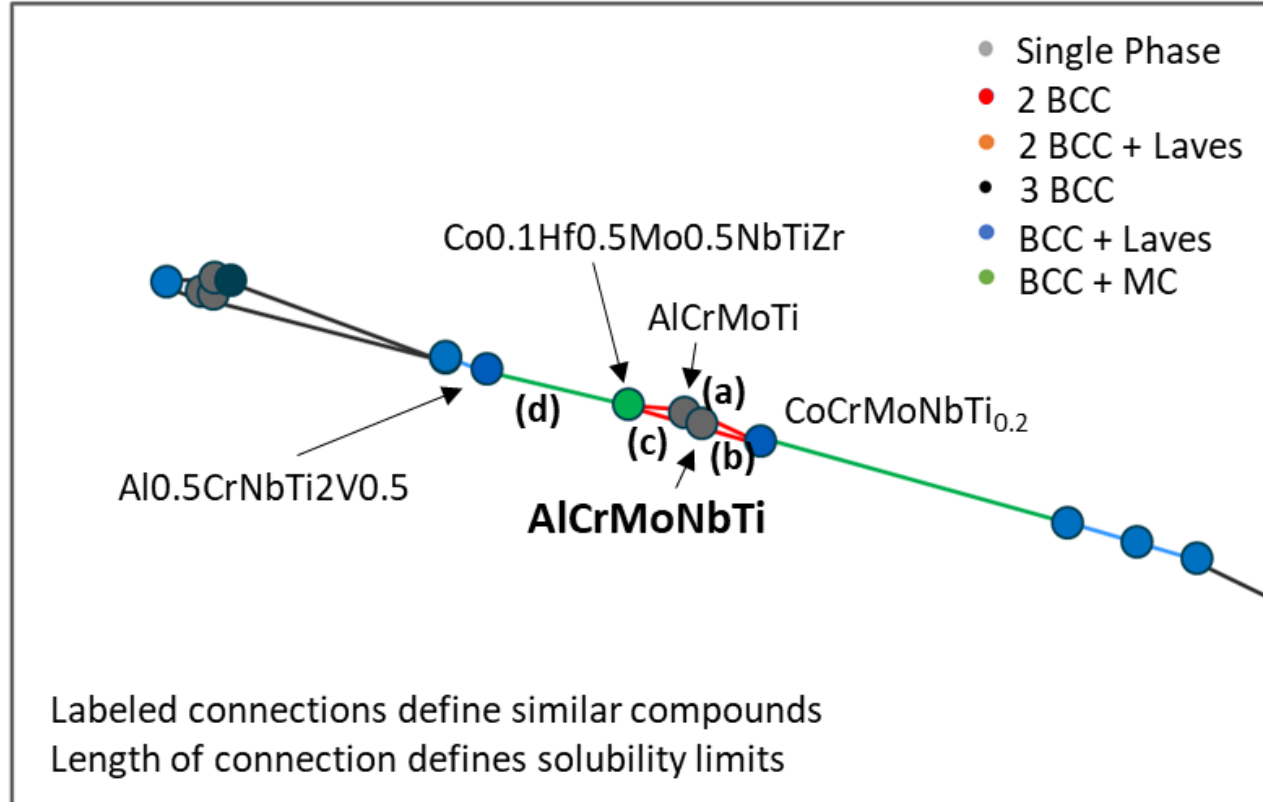
Selection of Features
of Principal Components

IsoMap :

Selection of Features
Distance Metric
of Nearest Neighbors

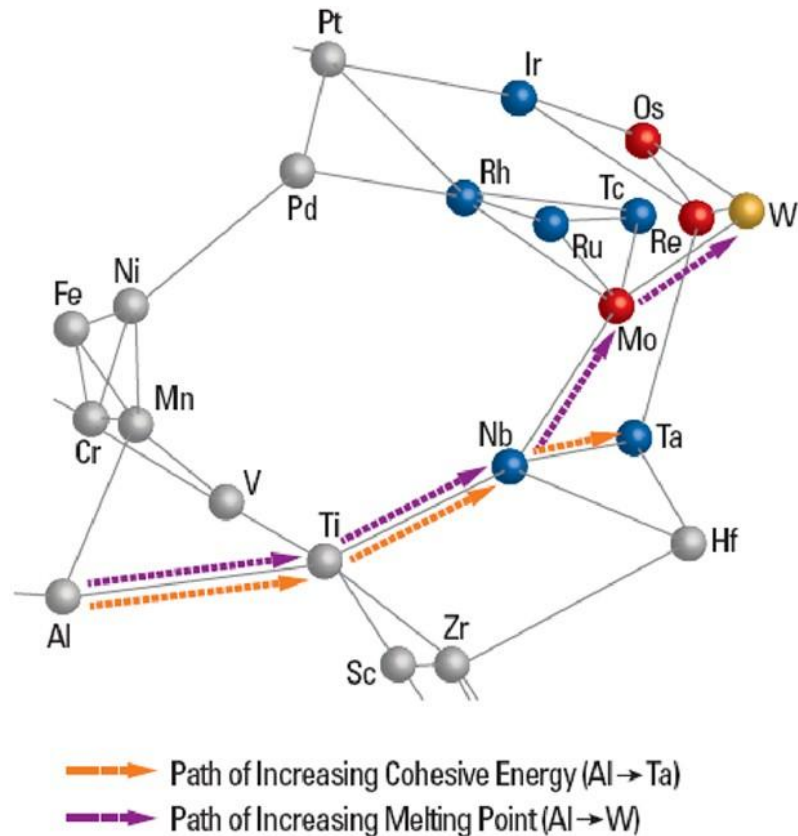


Capturing Phase Changes through Network



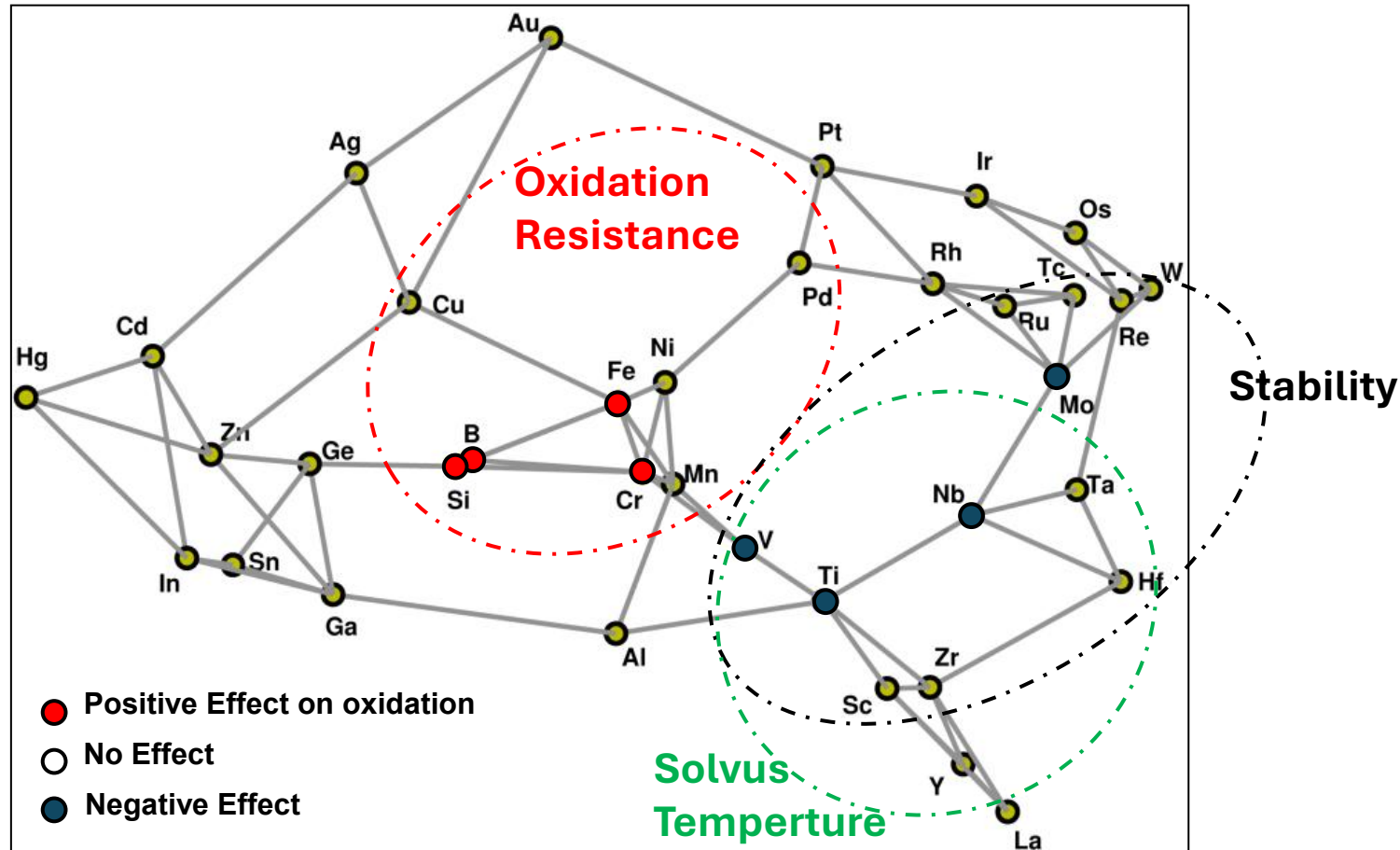
- The connection of points (alloys) defines similarity of compounds, across all of the data – can be considered akin to a tie-line
- Tracking changes across connections defines subtle differences which result in property changes.

Stability of Co_3Al



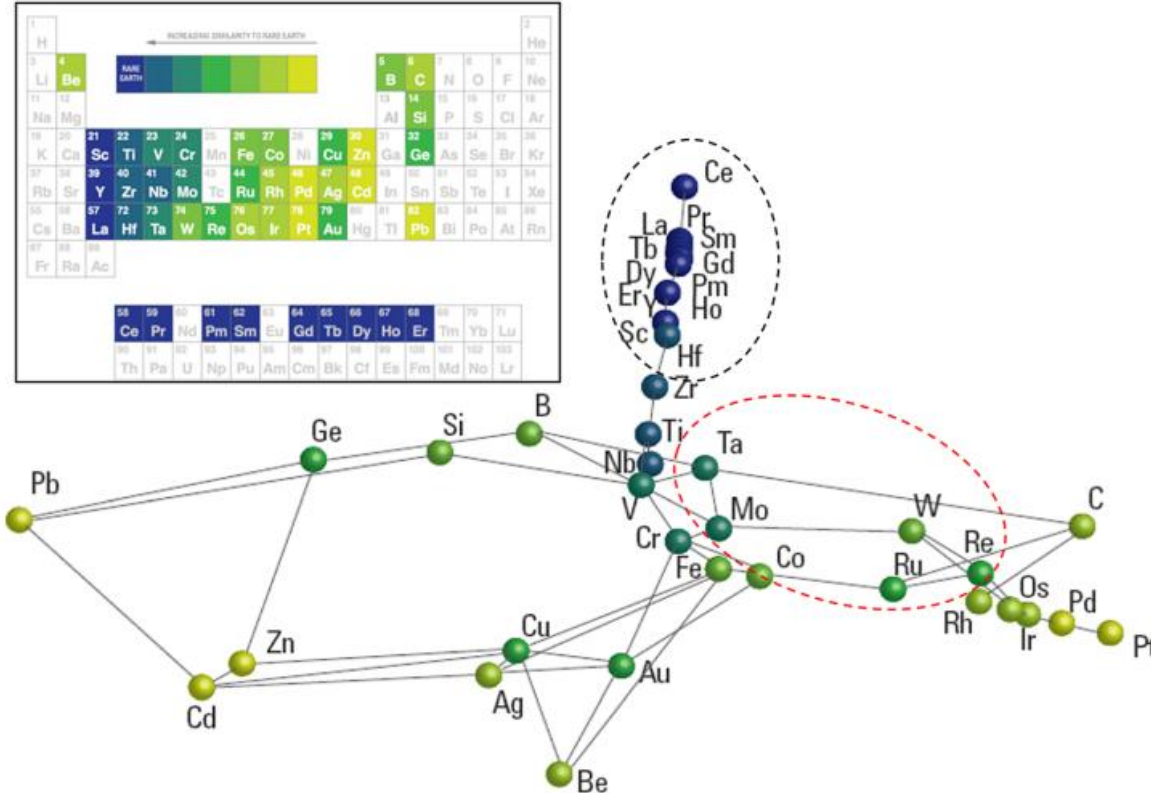
- Known that with the addition of W, $\text{Co}_3(\text{Al}, \text{W})$ is a stable intermetallic possessing high melting point, L1_2 ordered structure, appropriate lattice parameter to achieve coherency strains that enhance high temperature mechanical properties of cobalt alloys typical to nickel based superalloys.
- Ti is a key node on the network pathway for higher cohesive energy. V is a nearest neighbor with Ti.
- Mn and Cr are *not* a nearest neighbor with any node on the network pathway, suggesting that these are not expected to contribute to γ' stability.
- Experimental and computational analyses are consistent with this result.

Regions of Properties in Cobalt 'Periodic Table'



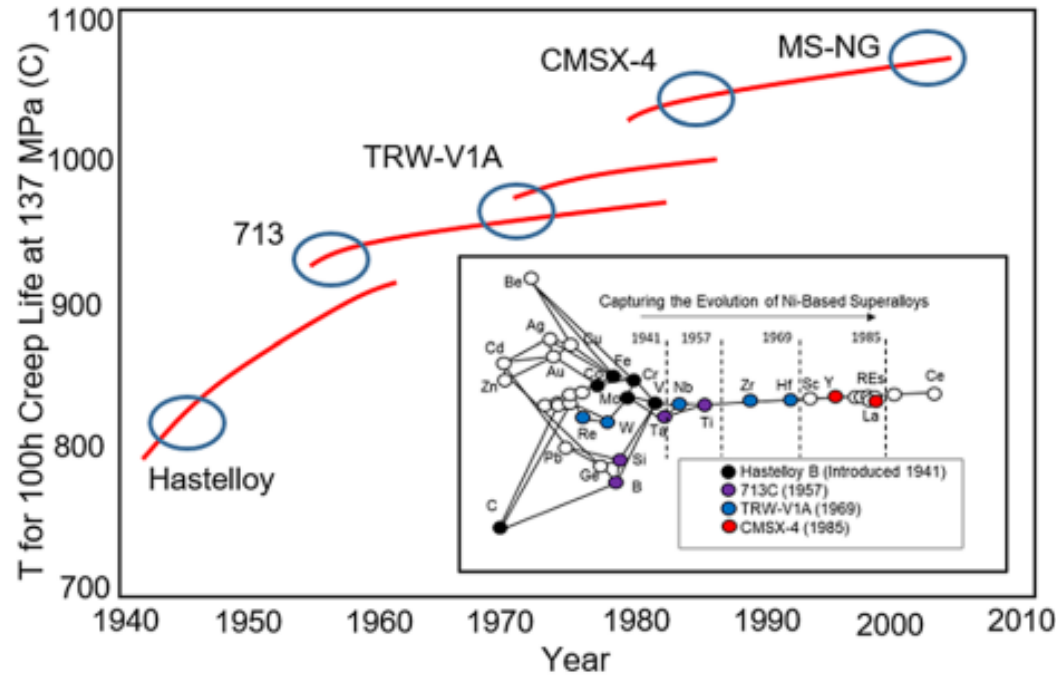
- The oxidation substitute region is defined based on the available data in literature. Nearest neighbor connections to this region have higher probability of beneficial effect (e.g. B, Cr, Fe, Mn, Ni, Si, Pd, and Cu)
- Can track the role of other properties and capture the impact of various properties simultaneously.

Replacement of RE elements in Ni Superalloys

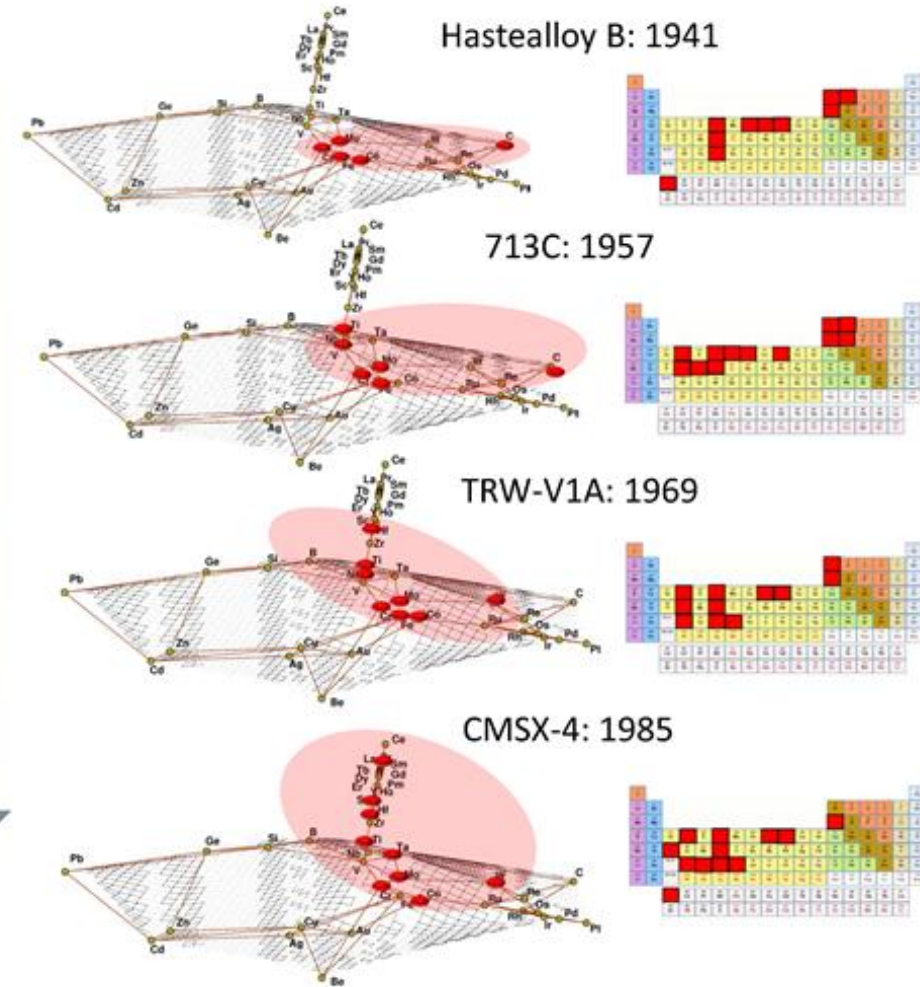


- Based on connections, density of nodes, and extracted design pathways, possible chemical substitutions are extracted, which are otherwise not obvious.
- We find Ti, Zr, and Hf sit on the rare earth pathway, while Ta, Nb, and V sit at the intersection of this pathway and the plane containing non-rare earth (RE) elements.
- These elements provide a logical starting point for rare earth substitutions by defining elements that connect rare earths with the rest of the periodic table.

Testing if Graph Captures Existing Knowledge



Time / Operating Temperature



Alloy Periodic Tables

Traditionally, the periodic table maps out the gaseous state property of elements. Now we have begun to propose a new series of periodic tables which instead map out specific alloy characteristics.

1	H	2	He
3	Li	4	Be
5	Na	6	Mg
7	K	8	Ca
9	Sc	10	Ti
11	V	12	Cr
13	Mn	14	Fe
15	Co	16	Ni
17	Cu	18	Zn
19	Ga	20	Ge
21	As	22	Se
23	Br	24	Kr
25	Rb	26	Sr
27	Y	28	Zr
29	Nb	30	Mo
31	Tc	32	Ru
33	Rh	34	Pd
35	Ag	36	Cd
37	In	38	Sn
39	Sb	40	Te
41	I	42	Xe
43	Cs	44	Ba
45	La	46	Hf
47	Ta	48	W
49	Re	50	Os
51	Ir	52	Pt
53	Au	54	Hg
55	Tl	56	Pb
57	Bi	58	Po
59	At	60	Rn
61	Fr	62	Ra
63	Ac	64	
65	Ce	66	Pr
67	Nd	68	Pm
69	Sm	70	Eu
71	Gd	72	Tb
73	Dy	74	Ho
75	Er	76	Tm
77	Yb	78	Lu
79	Th	80	Pa
81	U	82	Np
83	Pu	84	Am
85	Cm	86	Bk
87	Cf	88	Es
89	Fm	90	Md
91	No	92	Lr

Stability Design
(Co alloys)

1	H	2	He
3	Li	4	Be
5	Na	6	Mg
7	K	8	Ca
9	Sc	10	Ti
11	V	12	Cr
13	Mn	14	Fe
15	Co	16	Ni
17	Cu	18	Zn
19	Ga	20	Ge
21	As	22	Se
23	Br	24	Kr
25	Rb	26	Sr
27	Y	28	Zr
29	Nb	30	Mo
31	Tc	32	Ru
33	Rh	34	Pd
35	Ag	36	Cd
37	In	38	Sn
39	Sb	40	Te
41	I	42	Xe
43	Cs	44	Ba
45	La	46	Hf
47	Ta	48	W
49	Re	50	Os
51	Ir	52	Pt
53	Au	54	Hg
55	Tl	56	Pb
57	Bi	58	Po
59	At	60	Rn
61	Fr	62	Ra
63	Ac	64	
65	Ce	66	Pr
67	Nd	68	Pm
69	Sm	70	Eu
71	Gd	72	Tb
73	Dy	74	Ho
75	Er	76	Tm
77	Yb	78	Lu
79	Th	80	Pa
81	U	82	Np
83	Pu	84	Am
85	Cm	86	Bk
87	Cf	88	Es
89	Fm	90	Md
91	No	92	Lr

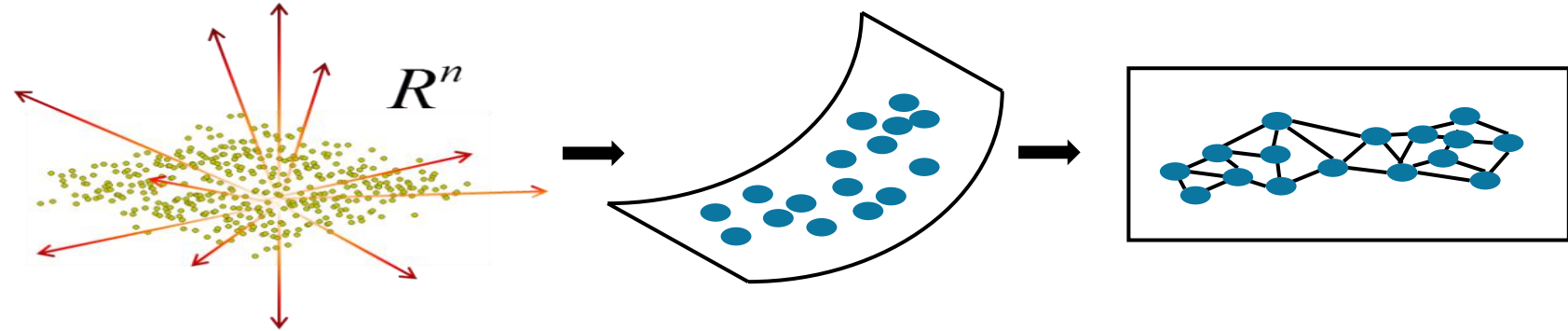
Critical Element
Substitution (Ni alloys)

1	H	2	He
3	Li	4	Be
5	Na	6	Mg
7	K	8	Ca
9	Sc	10	Ti
11	V	12	Cr
13	Mn	14	Fe
15	Co	16	Ni
17	Cu	18	Zn
19	Ga	20	Ge
21	As	22	Se
23	Br	24	Kr
25	Rb	26	Sr
27	Y	28	Zr
29	Nb	30	Mo
31	Tc	32	Ru
33	Rh	34	Pd
35	Ag	36	Cd
37	In	38	Sn
39	Sb	40	Te
41	I	42	Xe
43	Cs	44	Ba
45	La	46	Hf
47	Ta	48	W
49	Re	50	Os
51	Ir	52	Pt
53	Au	54	Hg
55	Tl	56	Pb
57	Bi	58	Po
59	At	60	Rn
61	Fr	62	Ra
63	Ac	64	
65	Ce	66	Pr
67	Nd	68	Pm
69	Sm	70	Eu
71	Gd	72	Tb
73	Dy	74	Ho
75	Er	76	Tm
77	Yb	78	Lu
79	Th	80	Pa
81	U	82	Np
83	Pu	84	Am
85	Cm	86	Bk
87	Cf	88	Es
89	Fm	90	Md
91	No	92	Lr

Microstructural Behavior
(Ti alloys)

Qualitative Predictions for Small, Sparse, and Uncertain Data

Building a Graph Network



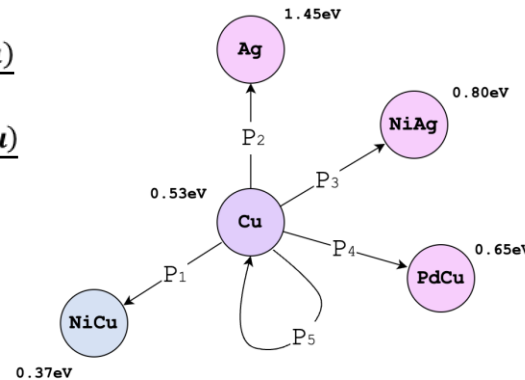
Random Walk

$$\text{Move with P2} = e^{\frac{-\Delta E}{kT}} = e^{\frac{-(E_{Ag}-E_{Cu})}{kT}}$$

$$\text{Move with P3} = e^{\frac{-\Delta E}{kT}} = e^{\frac{-(E_{NiAg}-E_{Cu})}{kT}}$$

$$\text{Move with P4} = e^{\frac{-\Delta E}{kT}} = e^{\frac{-(E_{PdCu}-E_{Cu})}{kT}}$$

$$P2+P3+P4+P5=1$$

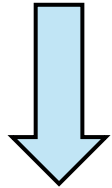


- Probability criteria (akin to ΔE) defined to best match the properties; *semi-supervised approach*
- The nodes (akin to energies) were labeled from the regression models; use of network and RW to mitigate error.

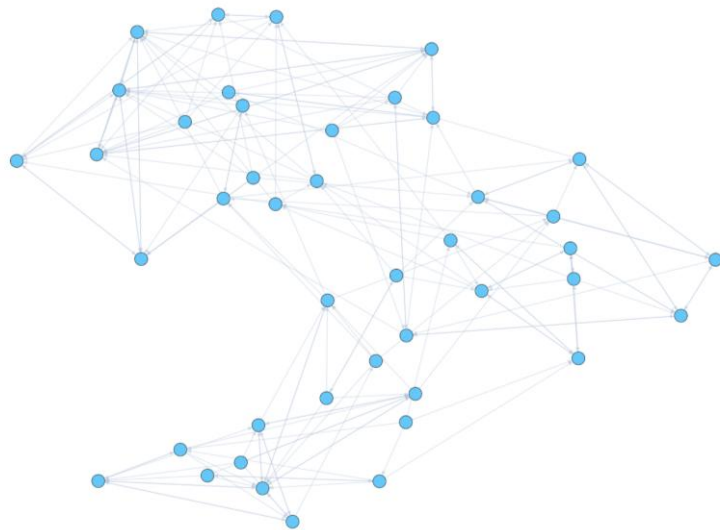
Qualitative Predictions for Small, Sparse, and Uncertain Data

Input

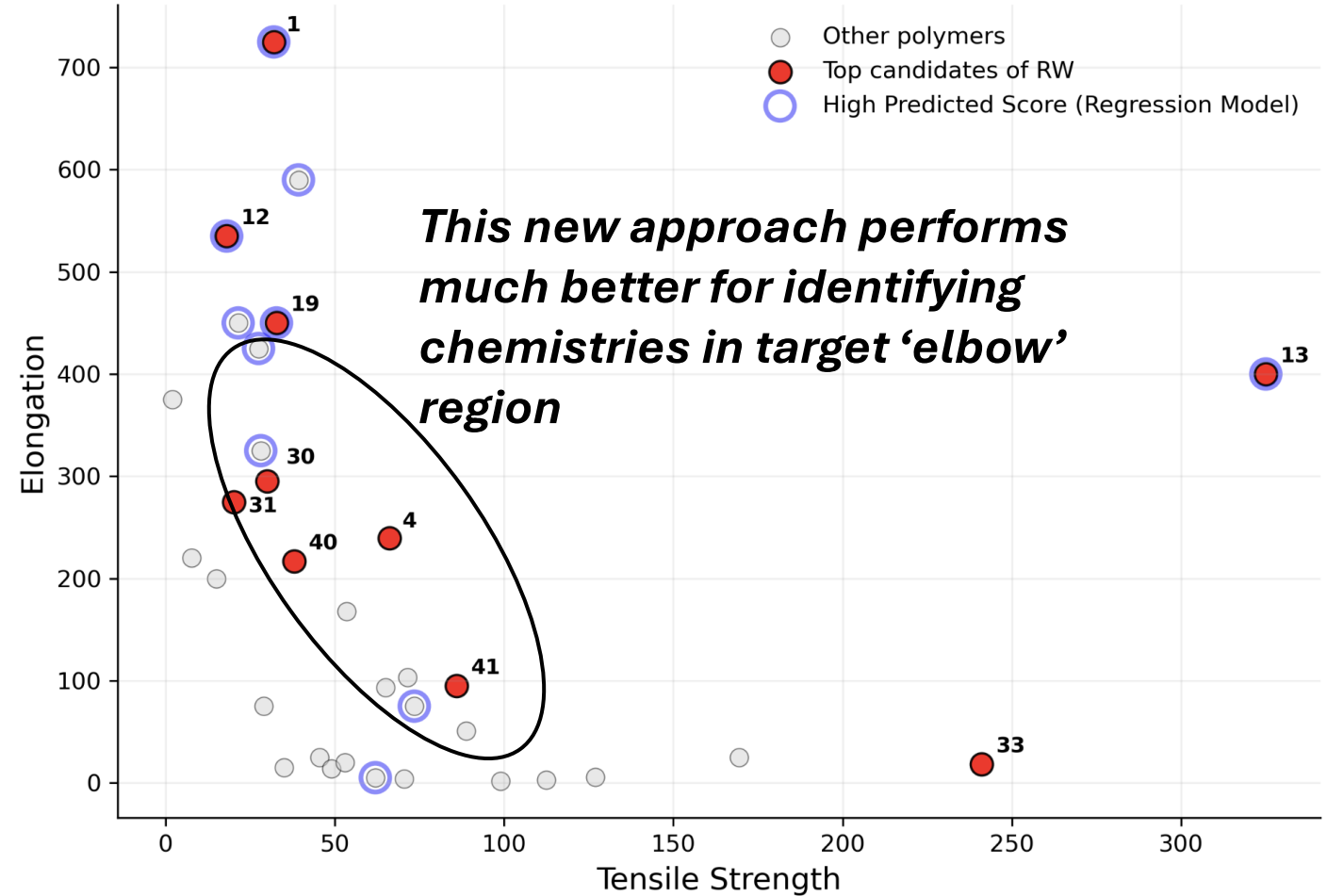
Topological Descriptors
Processing Parameters



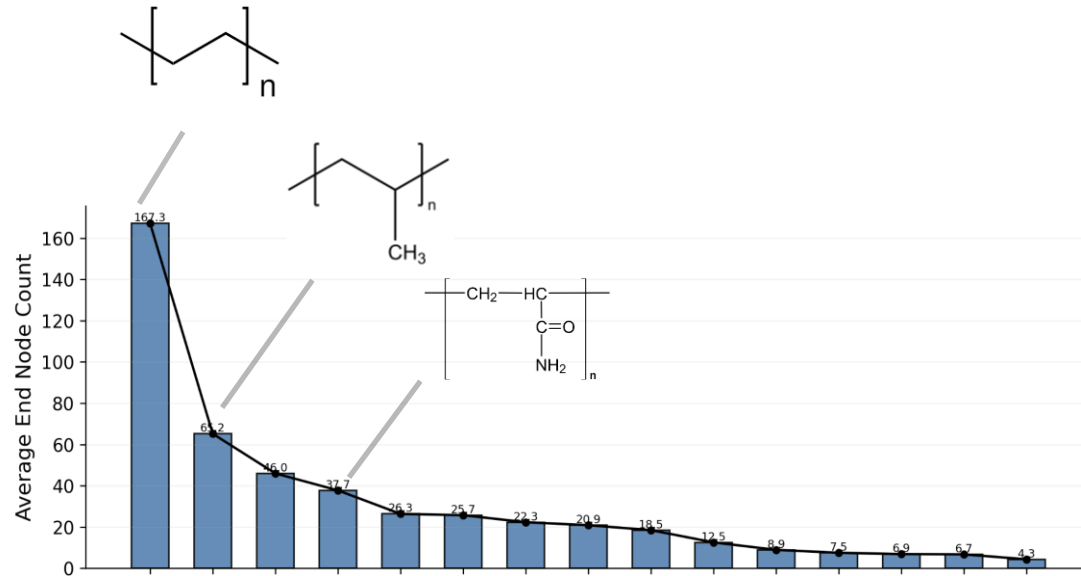
Graph Network



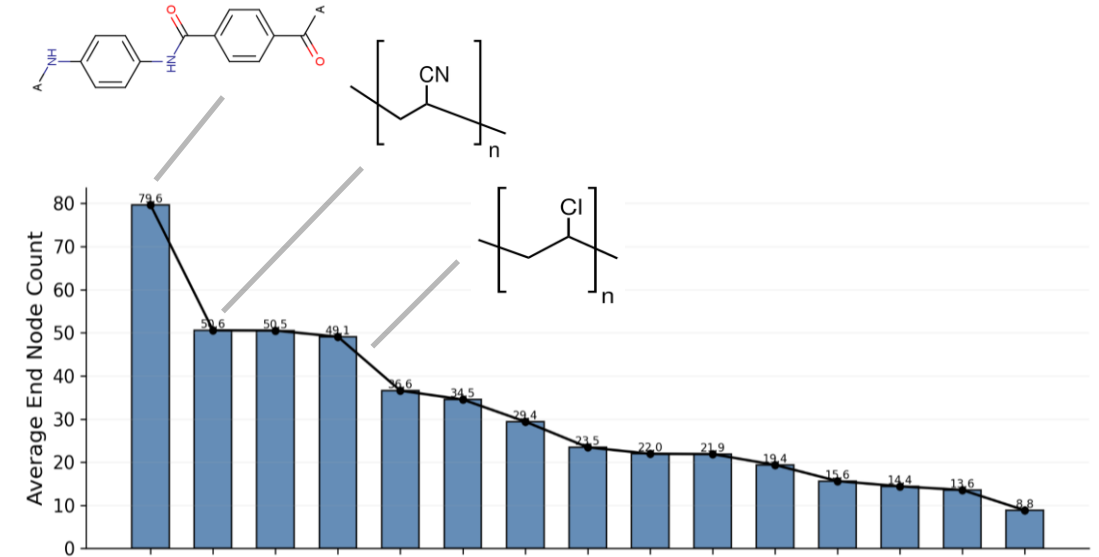
RW



Random Walk Selection for Mechanical Properties



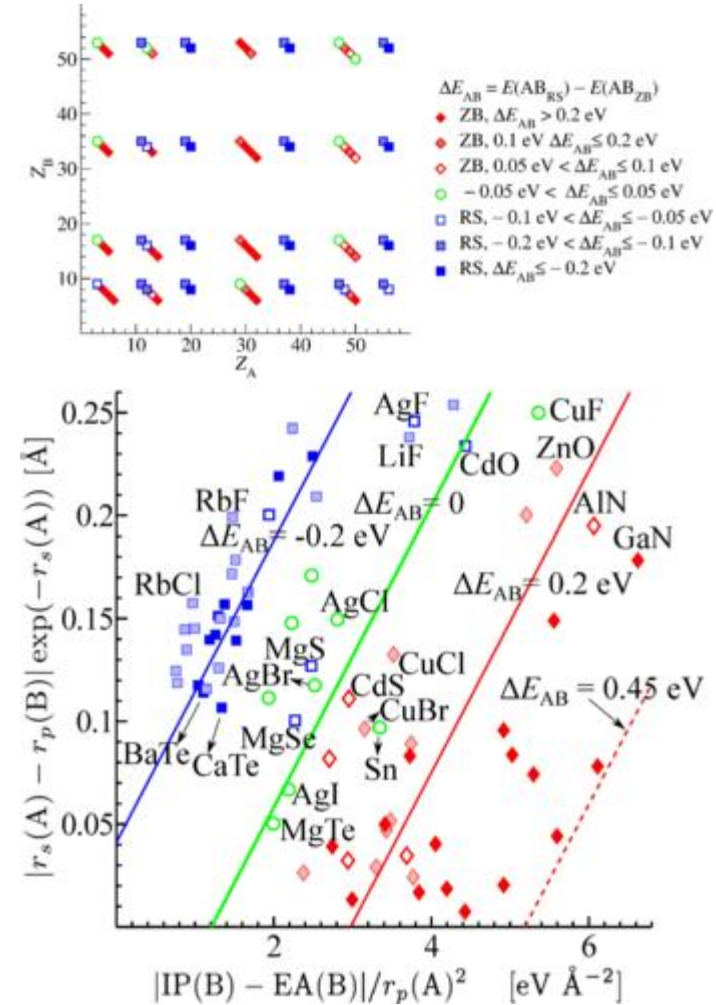
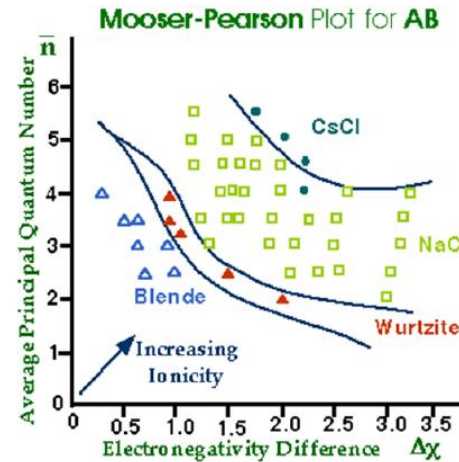
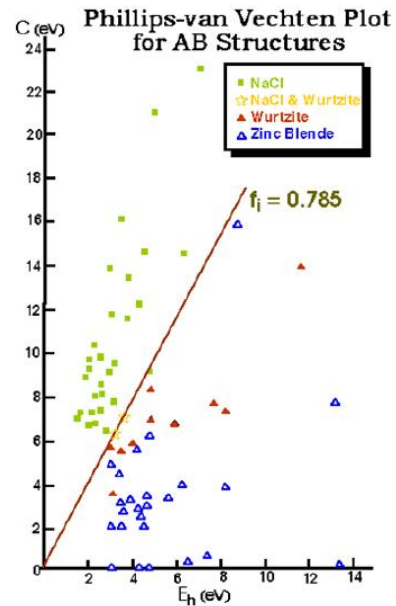
Elongation-Guided Random Walk



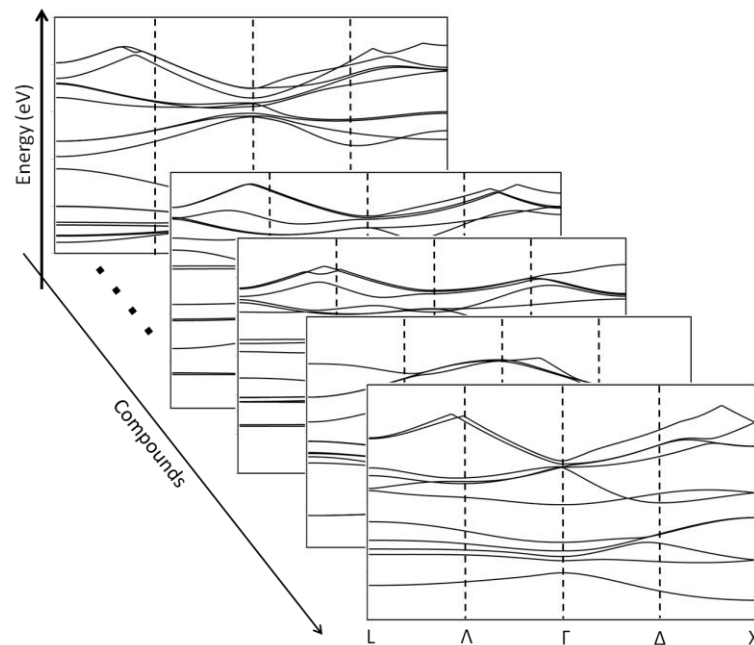
Tensile-Strength-Guided Random Walk

Elongation-guided walks favored flexible aliphatic structures, while strength-guided walks highlighted more rigid and polar structures.

One More Example : Octet Compounds



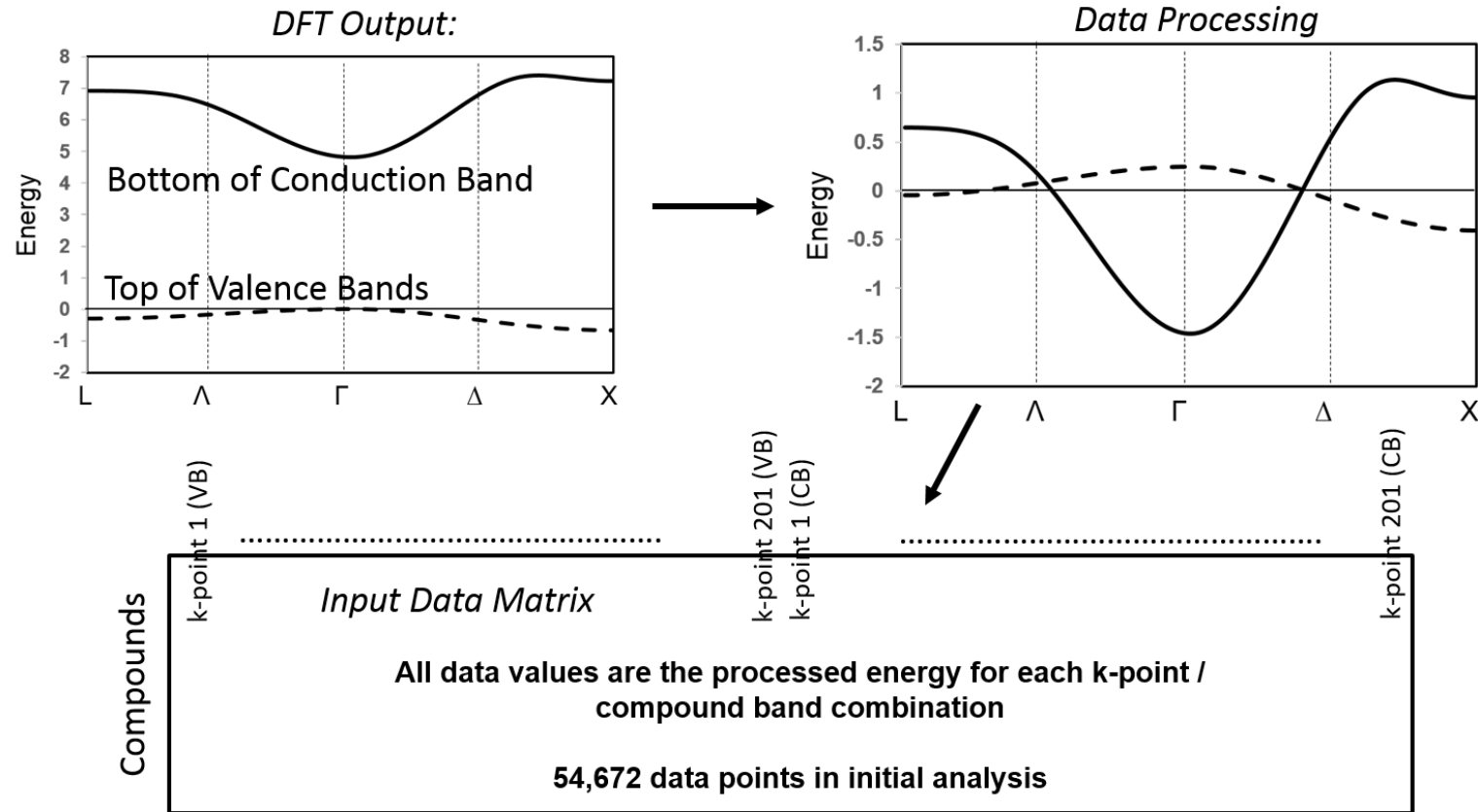
Band structures



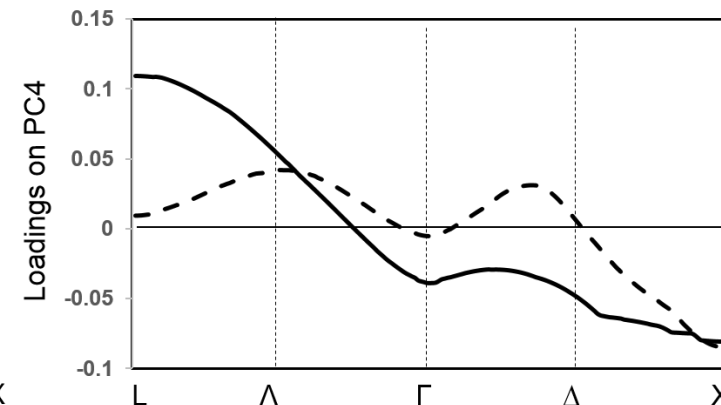
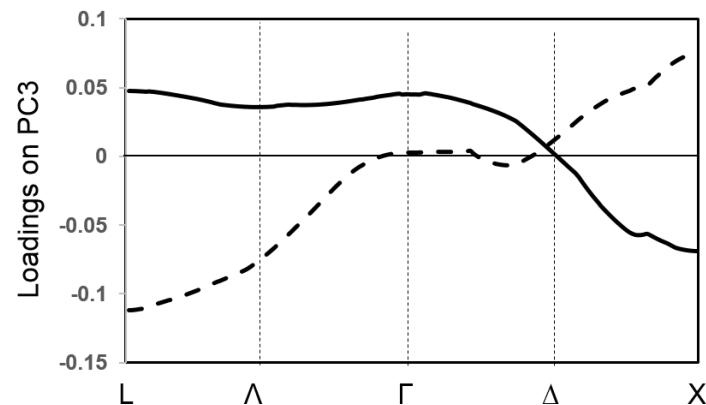
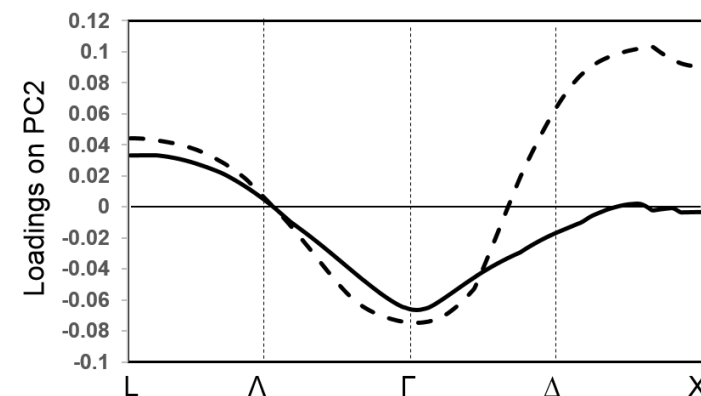
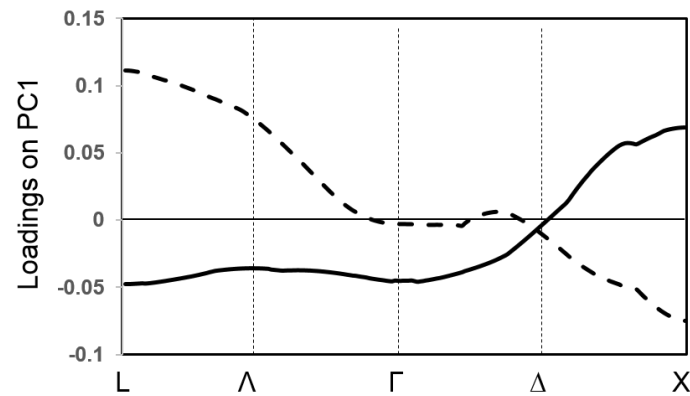
Data Matrix:

36 compounds, 201 k-points
> 100,000 data points

Screened all bands for all compounds --- top of valence band and bottom of conduction band are identified as sufficient for separating compounds



Loading Spectra

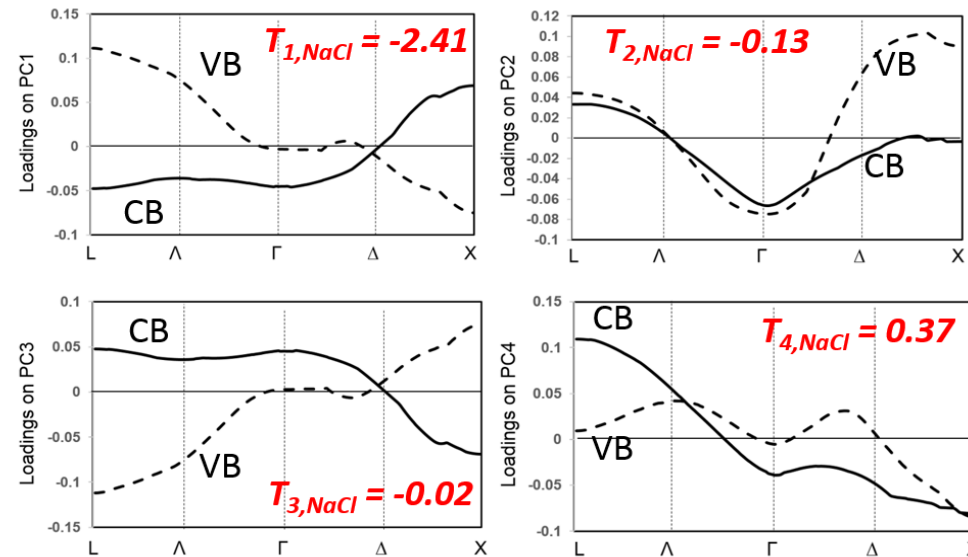


Identifies the patterns most distinguishing the compounds in terms of covariance.

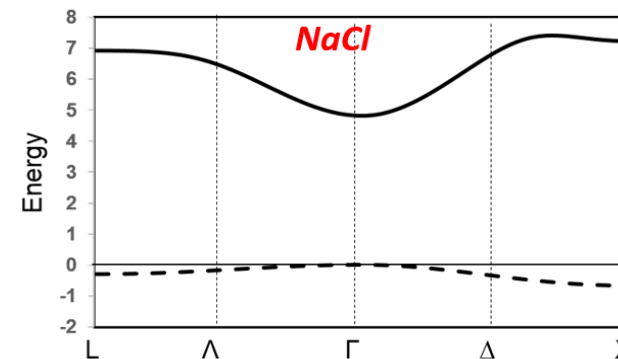
The first pattern (loadings on PC1) captures the largest amount of variance of the compounds.

Loading Spectra

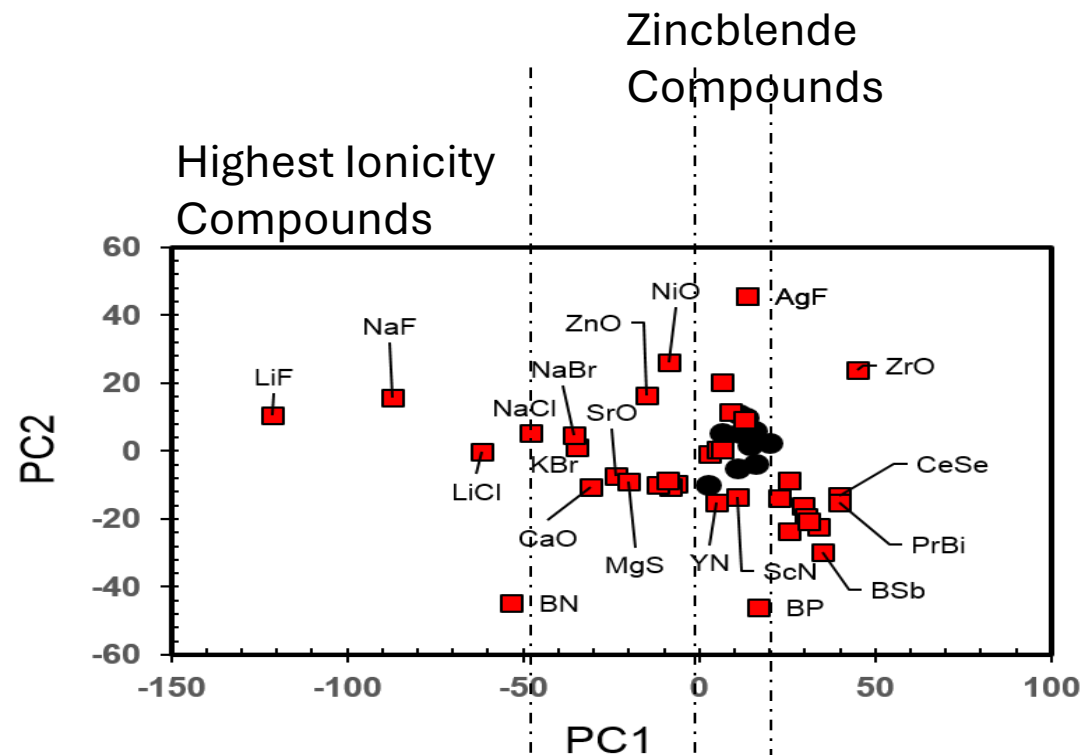
Contribution of the loadings spectra to the band structure. The two outputs from the analysis are the primary patterns differentiating the compounds and the weighting of the patterns on the total band structure for each compound. This example shows the deconvolution of the NaCl band structure.



$\Sigma T_{NaCl} * P^T$ ↓ and undoing pre-processing



Scores : Differentiating Crystal Structures



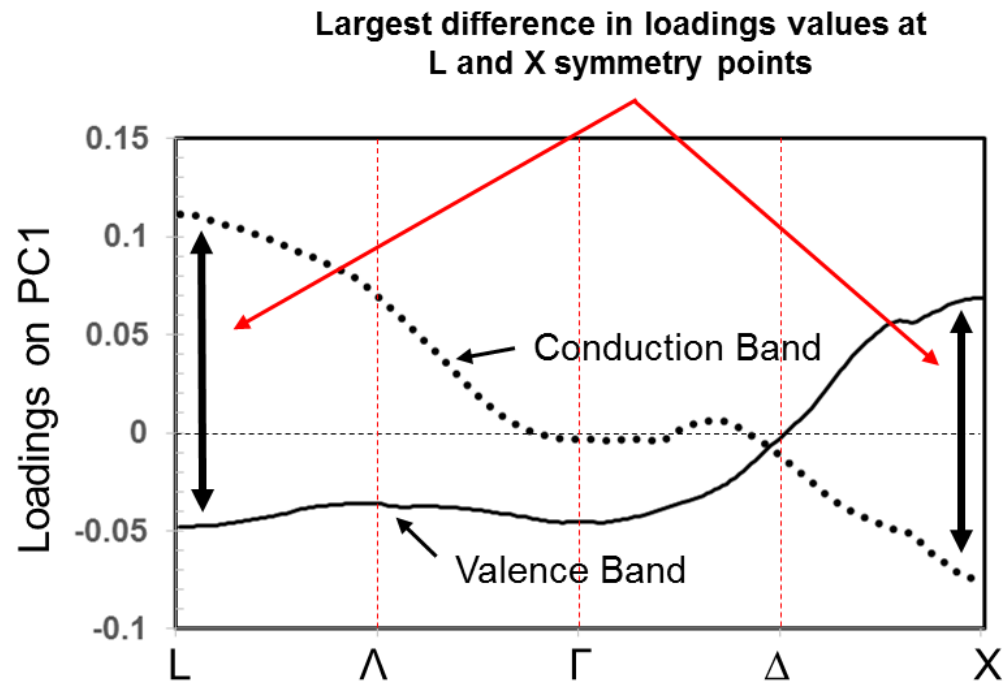
PC1 is capturing general trend in ionicity

This provides a mapping of the weighting of the signatures on the band structure for each compound --- 2D mapping of band structure spectra.

Following supervised approach, we find separation of compounds in PC1 – PC2 map. All zinc blende compounds are clustered together.

Highest ionicity compounds (LiF, NaF, LiCl, NaCl) have highest magnitude PC1 values. Zincblende compounds have PC1 value near zero.

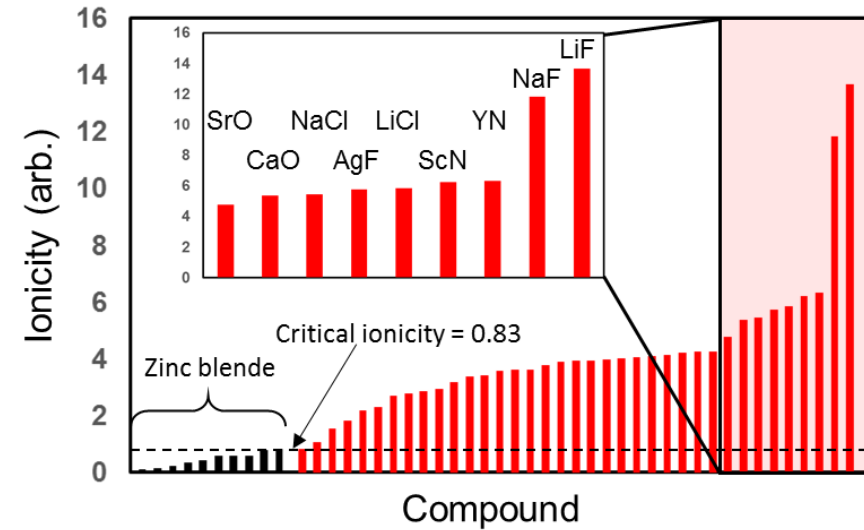
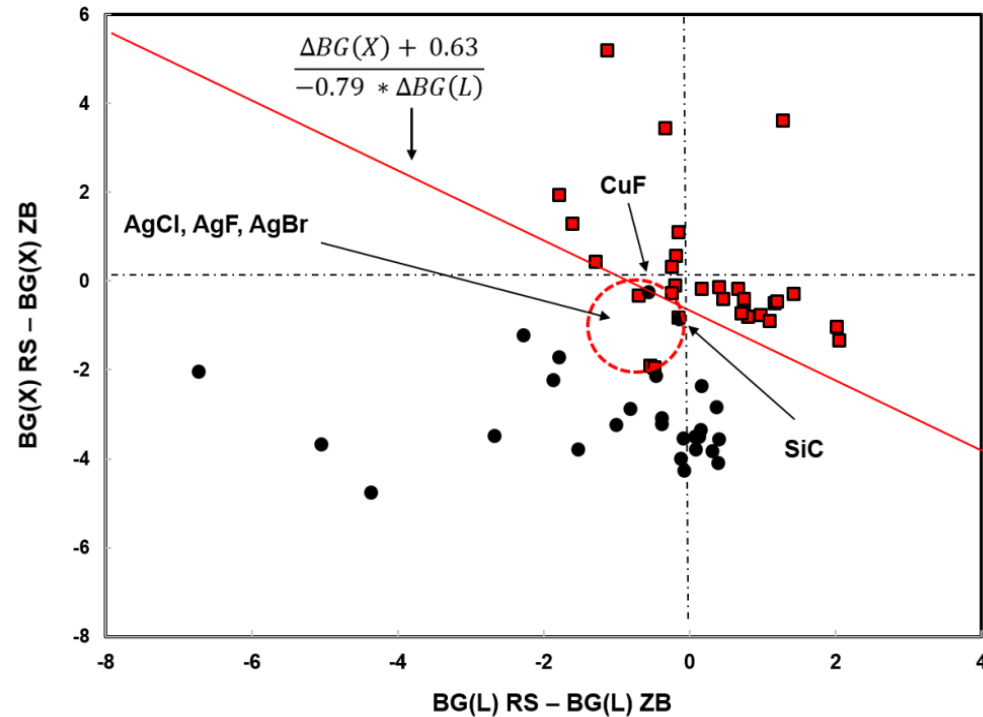
We identify PC1 as potentially capturing difference in ionicity. We look at the PC1 spectra to see what signatures dictate this.



The two primary signatures associated with the loadings spectra associated with PC1 are the difference at L and at X.

Bandgap at L (BG-L) and bandgap at X (BG-X) as the two critical descriptors capturing ionicity.

To evaluate this, we plot all the compounds based on these two parameters



Calculated band structure parameter, based on the informatics defined structure map, and with the parameter values calculated with SVM and a square cosine projection. The separation of RS versus ZB structures is very strong with this parameter (an improvement over P - vV), while the trend with ionicity is captured with LiF having the highest value and C having the lowest.

Summary

- Discussed unsupervised techniques, along with assumptions and biases, particularly PCA and IsoMap.
- Few case studies to demonstrate possible applications : development of new mappings, prediction of properties with small data, and identify key parameters in big data.
- Will now go through some simple code in R related with these topics.
- Next lecture will focus on supervised learning --- instead of inputting data and letting math determine output without assumed result, supervised learning assumes answer and attempts to find best fit to data. Will largely use electronic structure data and calculation outputs to design for material properties.

