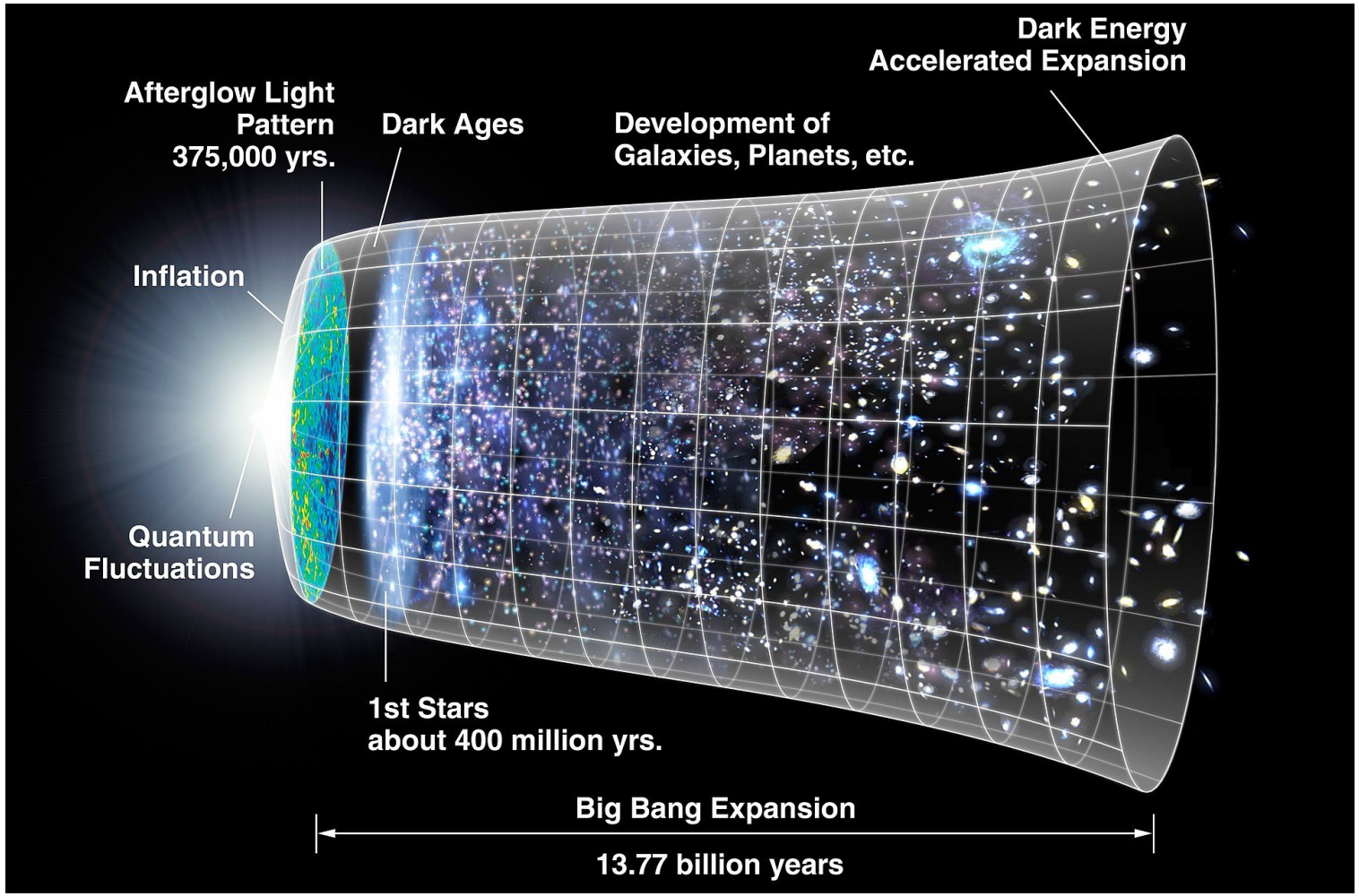
THE UNIVERSITY of TENNESSEE KNOXVILLE

Lecture 11: How Can We Describe Molecules and Materials

Instructor: Sergei V. Kalinin

From Observations to Physics



Wikipedia, NASA

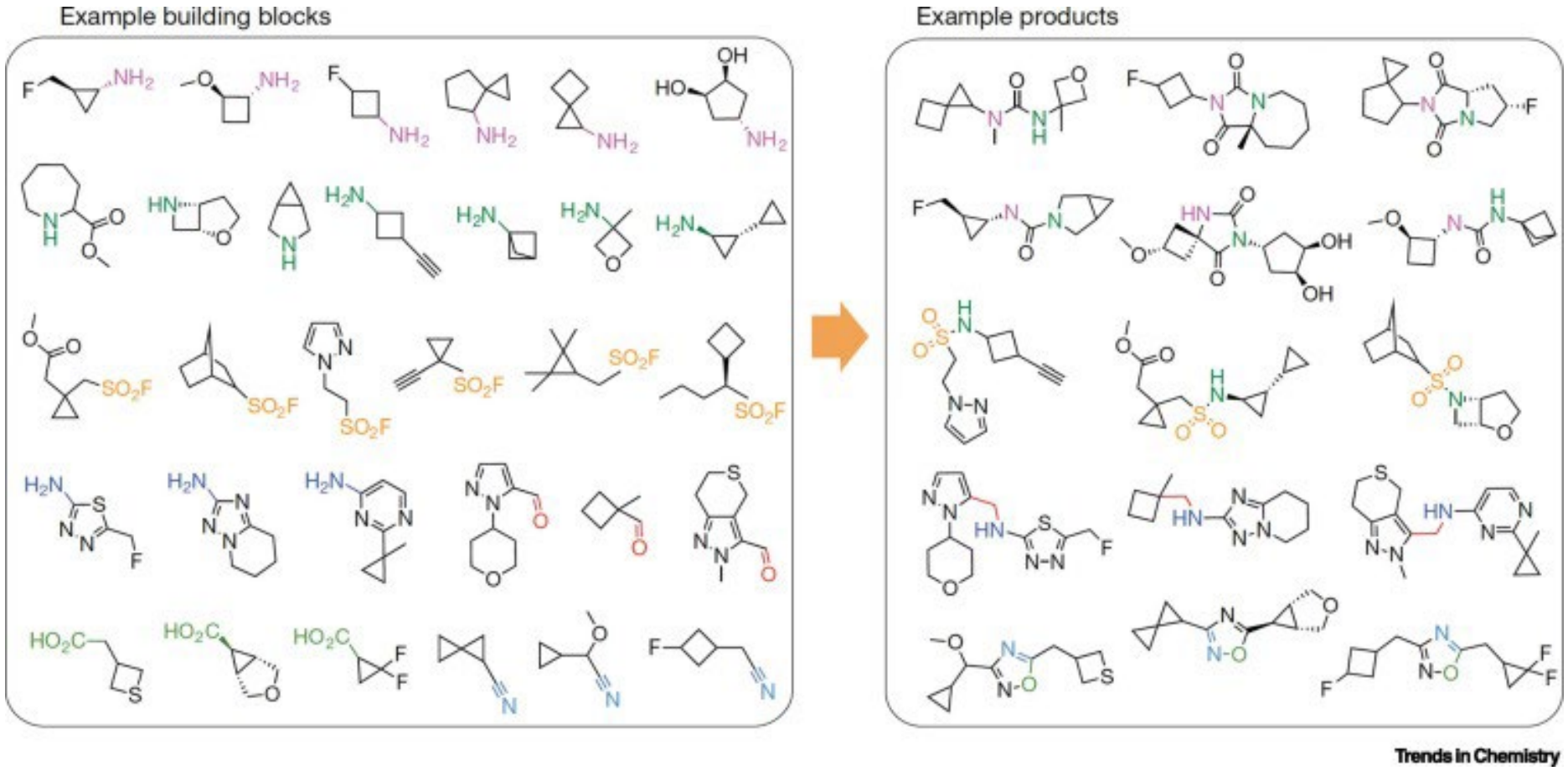
Levels of understanding

- Correlation
- Ptolemean model
- Copernican model
- Kepler Laws
- Newton laws
- Special relativity?

Levels of description:

- Point mass/Newton laws
- Inertia/rotations
- Lagrangian mechanics

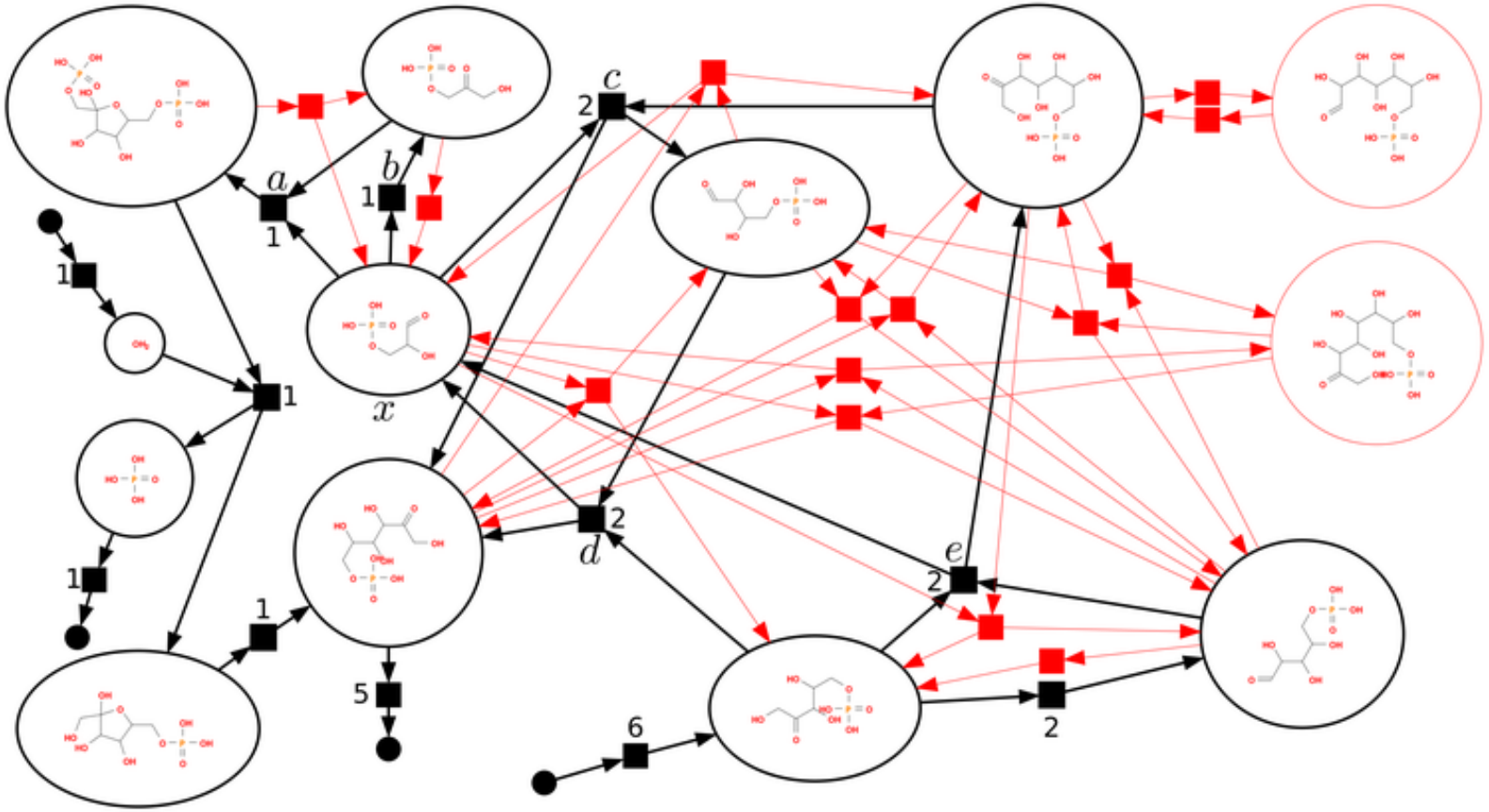
How many molecules are there?



A chemical space often referred to in cheminformatics is that of potential biologically active molecules. Its size is estimated to be in the order of 10^{60} molecules. The estimate restricts the chemical elements used to be C, H, O, N and S. It further makes the assumption of a maximum of 30 atoms to stay below 500 Daltons, allows for branching and a maximum of 4 rings and arrives at an estimate of 10^{63}.

https://www.cell.com/trends/chemistry/fulltext/S2589-5974%2820%2930288-4

Chemical reactions networks



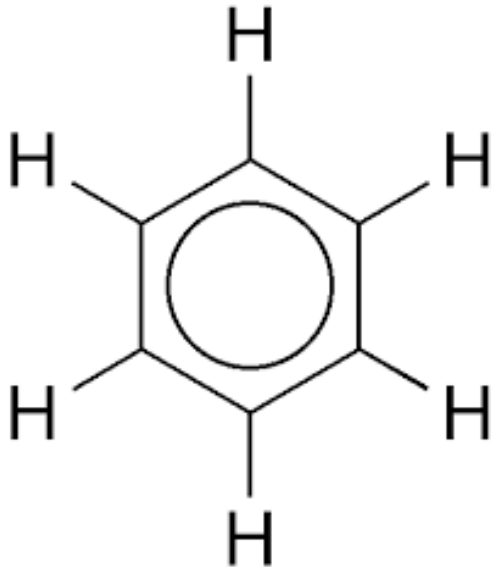
- Molecular property predictions: are they **likely** to be useful?
- Synthesizability scores: what would it **probably** take to make them
- Reaction network mining and retrosynthesis: can we identify **possible** synthetic pathways?
- Optimization of specific reaction conditions and pathways: myopic and non-myopic

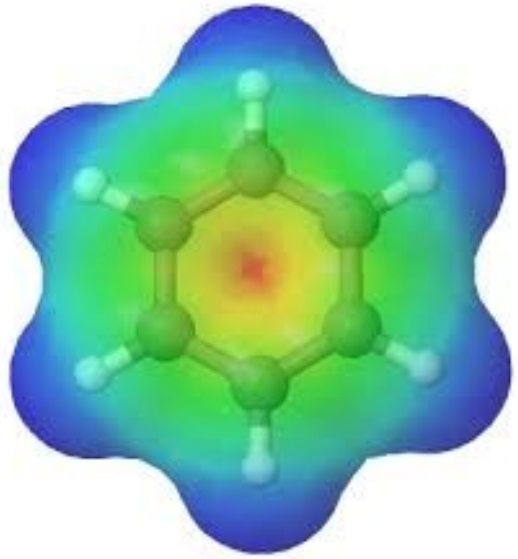
https://www.mis.mpg.de/stadler/research/chemical-reaction-networks.html

Descriptor or *fingerprint* are used, usually interchangeably, in chemical and materials informatics to indicate heuristically determined properties that are easier to compute than the quantities one ultimately wants to predict, but correlate strongly with them, facilitating the construction of transferable and accurate models.

https://pubs.acs.org/doi/10.1021/acs.chemrev.1c00021

Start simple: SMILES and SELFIES





SMILES (Simplified Molecular Input Line Entry System): A text notation of a molecule's 2D graph—atoms, bonds, branches () and ring closures 1,2,…—with most hydrogens implicit. Compact, human-readable, and ubiquitous in cheminformatics.

SELFIES (SELF-referencing Embedded Strings): A tokenized molecular string where every sequence of tokens decodes to a valid molecule. Tokens are bracketed (e.g., [C], [O], [Branch1_1], [Ring1]) and a built-in grammar enforces valence, making it robust for generative ML.

Molecule	SMILES	SELFIES (illustrative)
Propane	`CCC`	`[C][C][C]`
Methylcyclohexane	`CC1CCCCC1`	`[C][C][Branch1_1][C][C][C][C][C][Ring1][Ring1]`

However, life can be more complex

- We can have composition
- Structural chemical formulae
- Some measured properties
- Atomic coordinates from simulations or measurements
- Results of DFT/MD calculations of electronic properties
- Characterization: NMR, Raman/IR, etc.

1. **Topology / Graph (2D)** – from the molecular graph only
   Examples: Morgan/ECFP, MACCS, AtomPair, scaffolds (Bemis–Murcko), ring counts

1. **3D Geometry & Shape (needs conformer)** – captures spatial arrangement
   Examples: Coulomb Matrix (eigs), Bag-of-Bonds, shape/PMI/WHIM, surface area/volume

2. **Local Environment / Many-Body (3D)** – rotationally invariant neighborhoods
   Examples: SOAP, ACSF, MBTR, SLATM (often pooled over atoms)

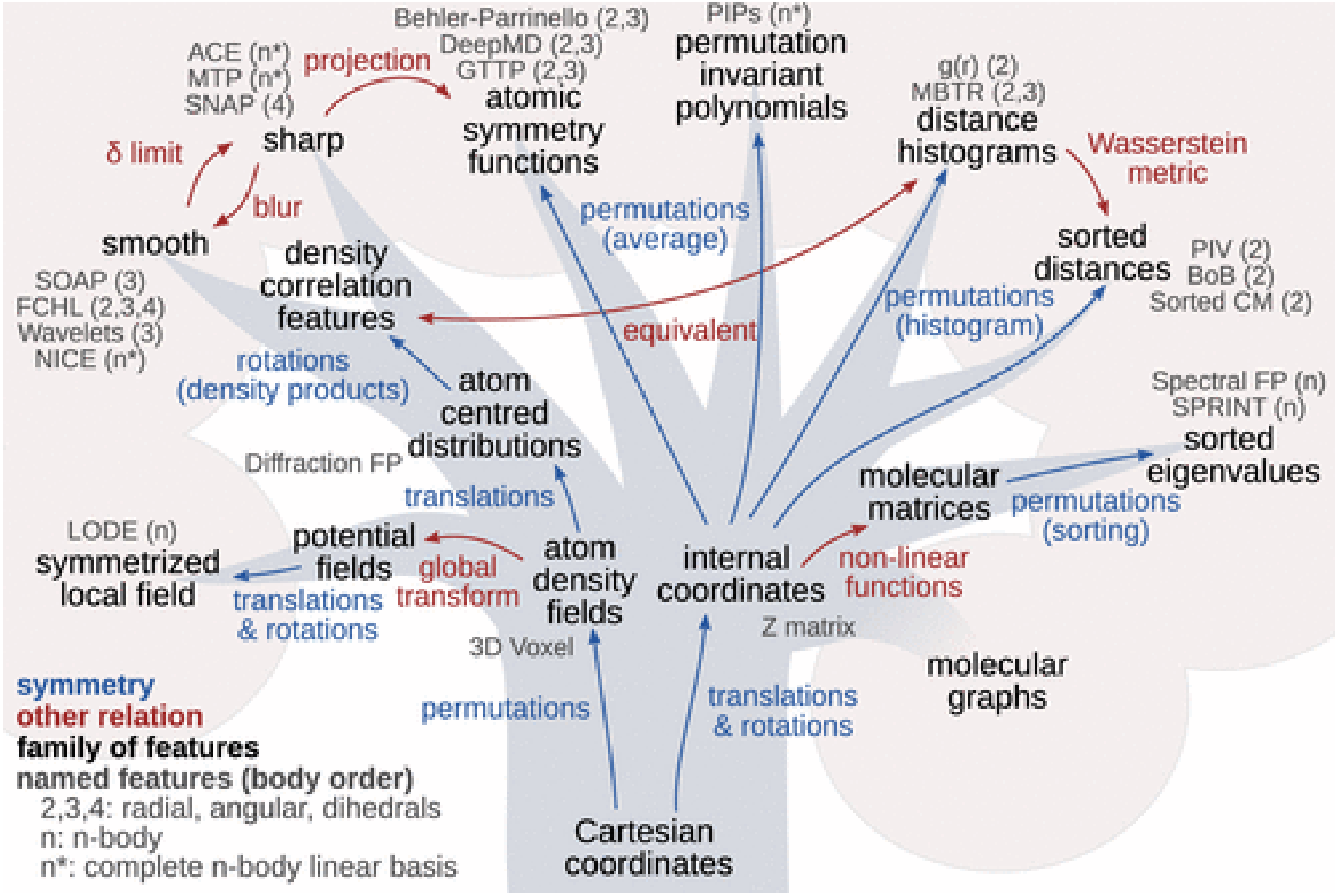
3. **Electronic / QM-derived** – from quantum calculations
   Examples: HOMO/LUMO/gap, dipole, polarizability, partial charges (Mulliken/NBO)

4. **Physicochemical Scalars (2D)** – simple, interpretable properties
   Examples: MW, logP, TPSA, HBD/HBA, rotatable bonds, fraction sp3, aromatic fraction

5. **Pharmacophore & Interaction Fields (2D/3D)** – feature patterns for binding
   Examples: HBD/HBA/AROM/LIPO maps, 3D pharmacophore distances/triads, ESP grids

6. **Learned Representations (data-driven)** – no hand-crafted features
   Examples: GNN embeddings (SchNet/DimeNet/MPNN/MACE), SMILES Transformers (ChemBERTa).

How do we describe molecules in theory?



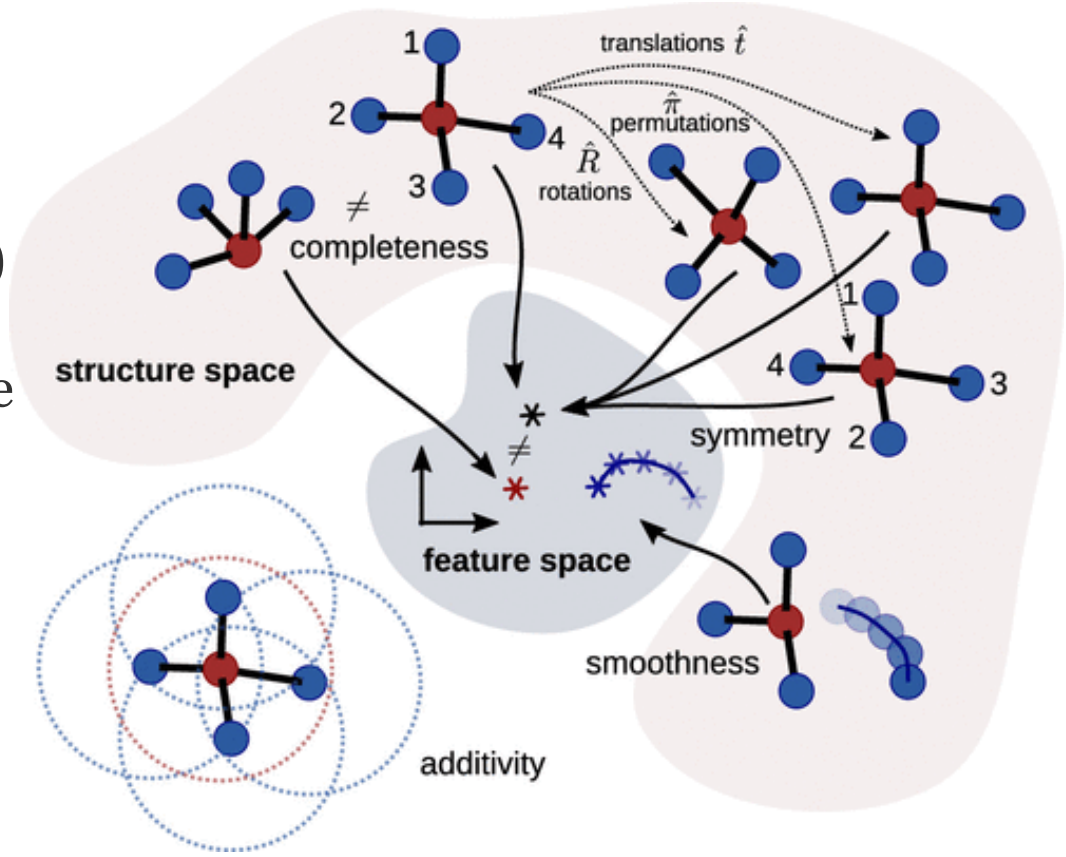
https://pubs.acs.org/doi/10.1021/acs.chemrev.1c00021

What are the requirements for descriptors?

(i) Invariance to transformations that preserve the predicted property, including (a) changes in atom indexing (input order, permutations of like atoms), and often (b) translations, (c) rotations, and (d) reflections. Predicting tensorial properties requires (e) covariance (equivariance) with rotations. Dependence of the property on a global frame of reference, for example, due to the presence of a non-isotropic external field, can affect variance requirements.

(ii) Uniqueness - variance against all transformations that change the predicted property: Two systems that differ in property should be mapped to different representations.

(iii)(a) Continuity, and ideally (b) differentiability, with respect to atomic coordinates.



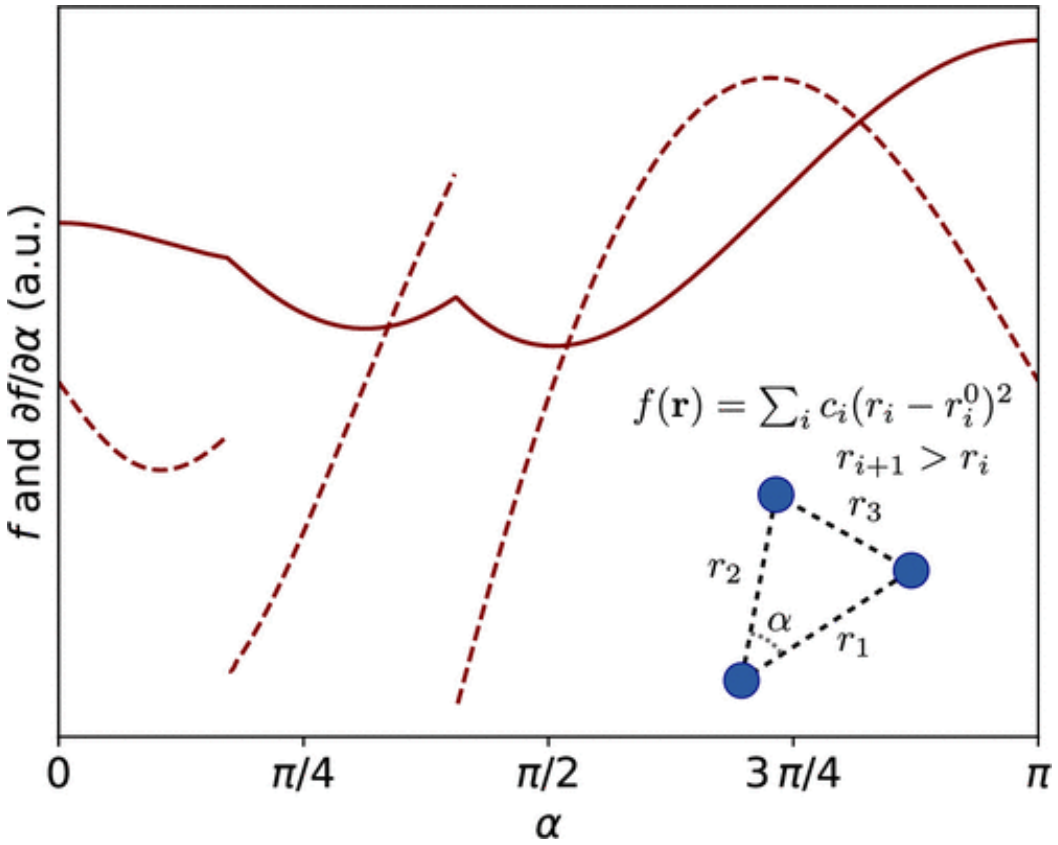
https://pubs.acs.org/doi/10.1021/acs.chemrev.1c00021

What are the requirements for descriptors?

(iv) Computational efficiency relative to the reference simulations.

(v) Structure of representations and the resulting data distribution should be suitable for regression.

(vi) Generality, in the sense of being able to encode any atomistic system. While current representations handle finite and periodic systems, less work was done on charged systems, excited states, continuous spin systems, isotopes, and systems subjected to external fields.



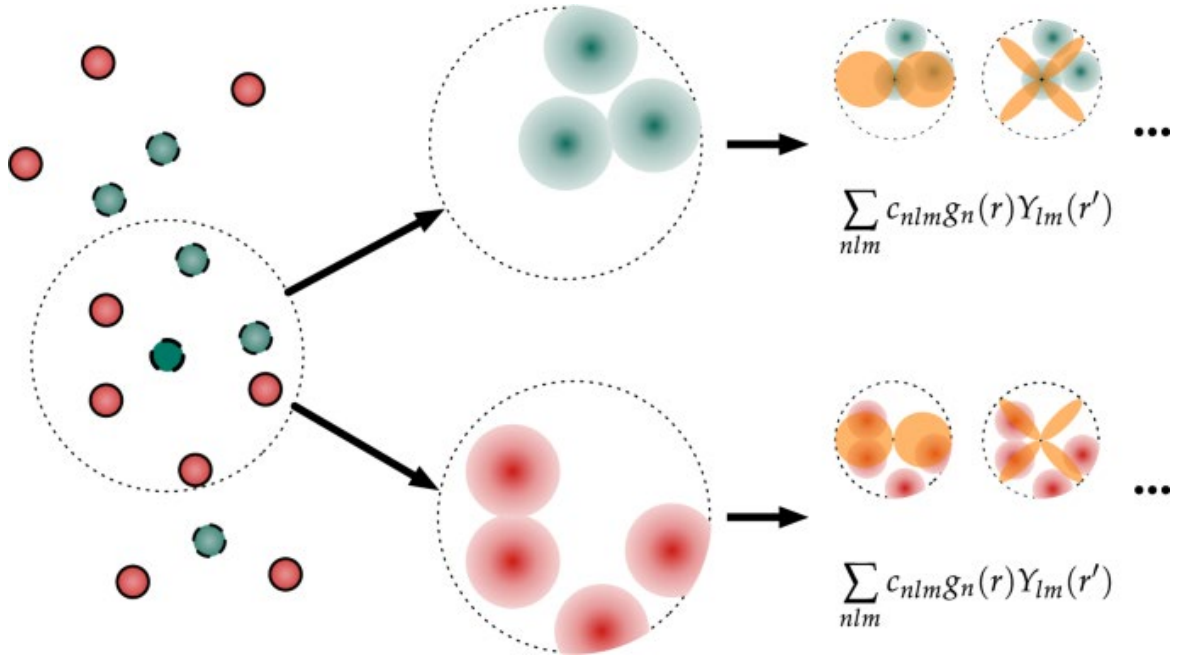
https://pubs.acs.org/doi/10.1021/acs.chemrev.1c00021

https://www.nature.com/articles/s41524-022-00721-x

Practical Tools

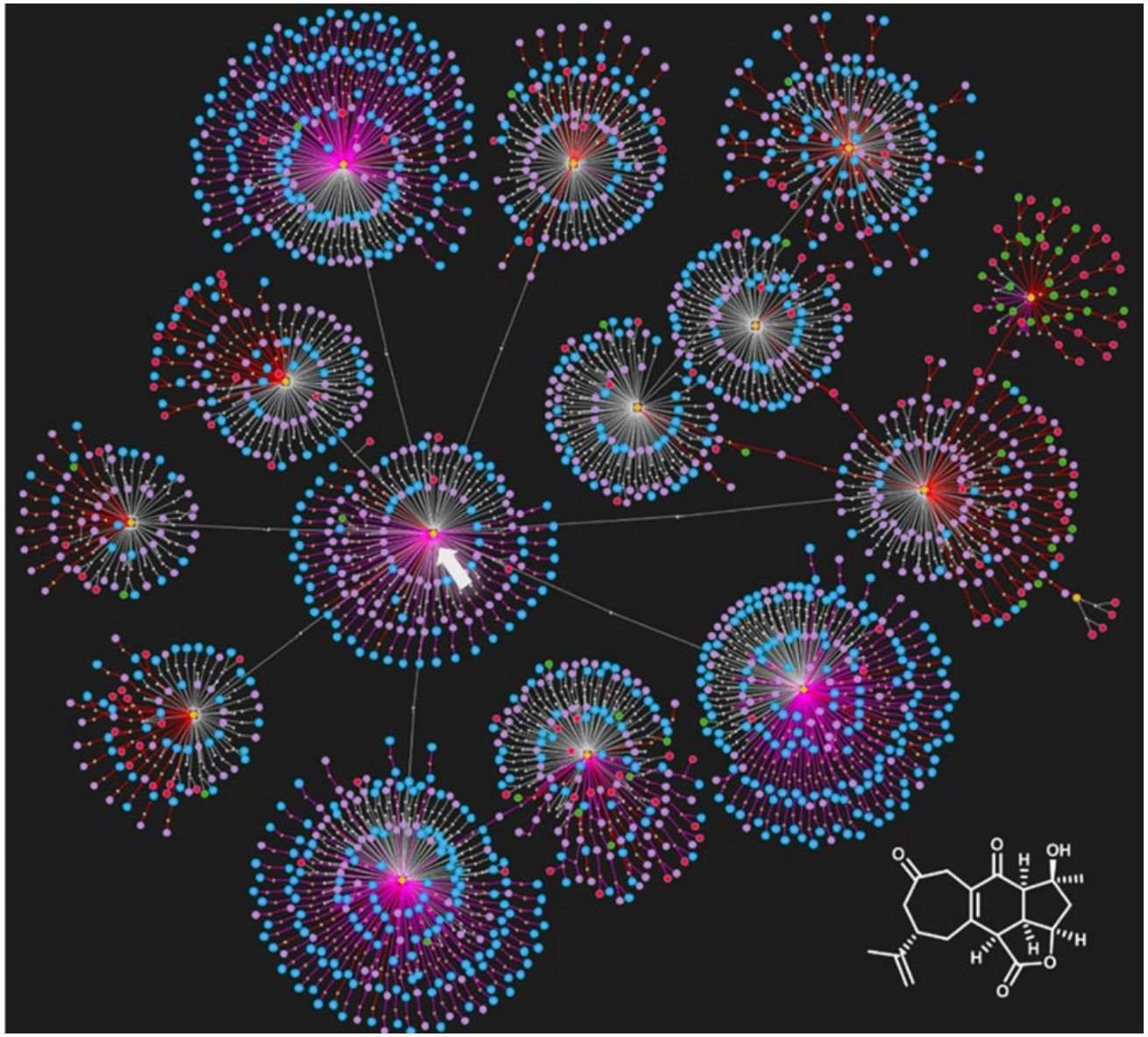
Tools:

1. **RDKit** (2D/3D basics, fingerprints, descriptors)
2. **DScribe** (MBTR, SOAP, ACSF, SLATM)
3. **Psi4/ORCA** (QM properties → features)
4. **DeepChem / PyG / HuggingFace** (learned embeddings)



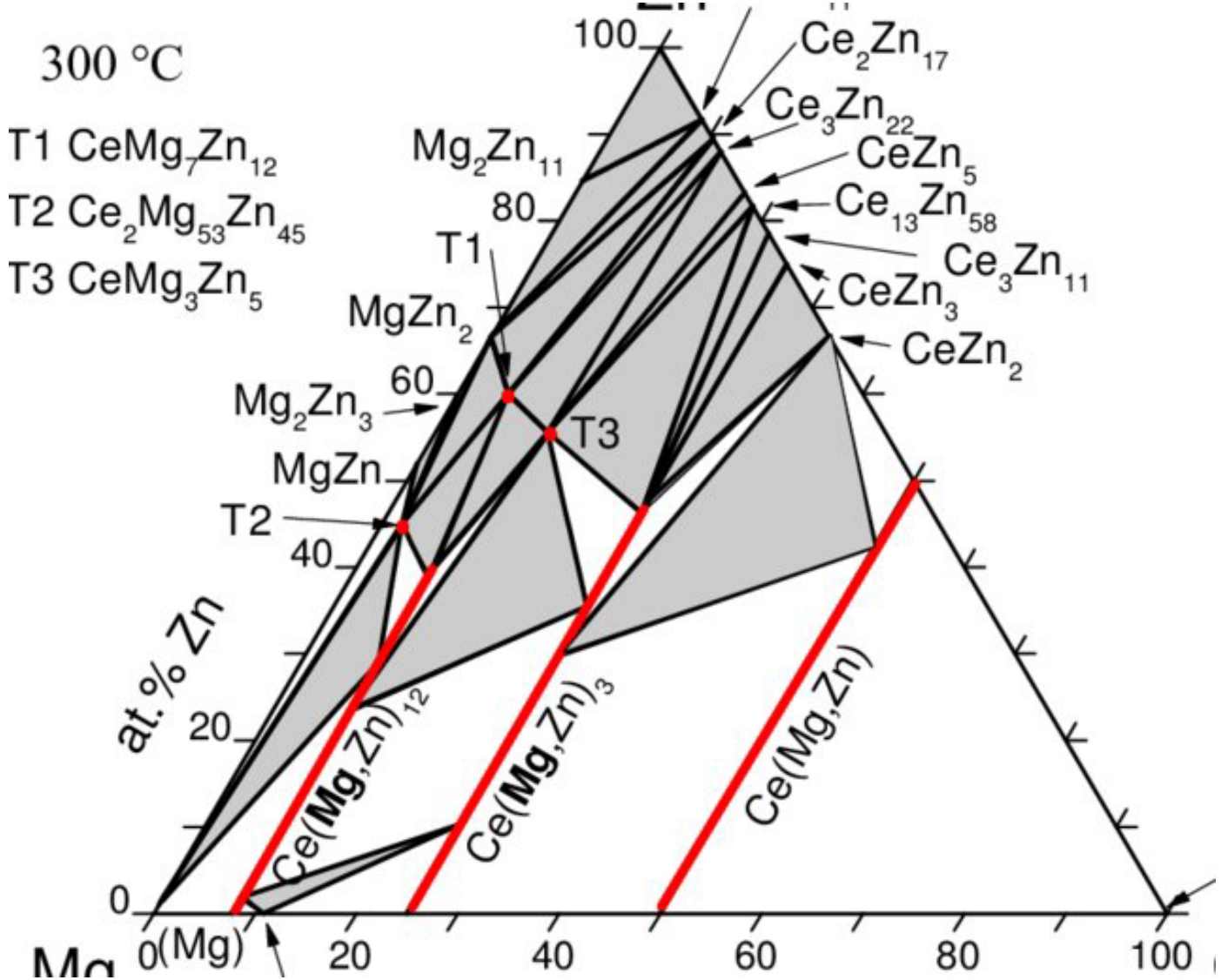
https://www.nature.com/articles/s41524-022-00721-x

What about synthesis planning?



1. Navigation in the space of known reactions (e.g. Chematica)
2. Optimization of single reactions
3. Retrosynthesis and ML based reaction planning

Inorganic Materials



Atomic Level

- Composition
- Structure
- Properties
- Phase transitions

Real Material

- Microstructure
- Macroscopic properties

Tasks for Inorganic Materials

1. **Structure & stability**
   - Structure from composition (prototype/polymorph prediction)
   - Stability ranking within a structural family; convex-hull distance / decomposition
   - Phase diagram construction (T–p–x), polymorph transitions

2. **Property prediction**
   - Electronic: band gap, DOS features, effective masses, Seebeck
   - Ionic/thermal: ionic conductivity, diffusion barriers, κ
   - Mechanical: elastic moduli, hardness, fracture toughness
   - Functional: dielectric/piezo/ferro/ferri/TC, superconducting Tc, catalytic activity

3. **Inverse design / multi-objective search**
   - Generate candidates meeting targets (e.g., wide-gap + high κ, Pareto fronts)
   - Active learning/BO for closed-loop experiments

4. **Defects/surfaces/interfaces**
   - Defect energetics & charge states, dopability, carrier concentration
   - Surface energies, adsorption energies, reaction energetics (Sabatier analysis)
   - Interface stability, strain engineering outcomes

Descriptors for Inorganic Materials

1. **Composition-only (no structure)**
   a. Element fractions, valence electron counts, Magpie-style stats (EN, radii, row/column)
   b. Oxidation-state heuristics, charge balance, bond-valence sums (BVS)
   c. Learned element embeddings (mat2vec, ElemNet-style)

2. **Crystal geometry & symmetry**
   a. Space group, Wyckoff positions, lattice metrics (a,b,c,α,β,γ), density, packing fraction
   b. Coordination numbers, bond-length/angle distributions, Voronoi volumes
   c. Radial/angle distribution functions (RDF/ADF), pair/bond statistics

3. **Local environment (rotationally invariant)**
   a. SOAP, ACSF/Behler–Parrinello, bispectrum/SO(3) power spectrum
   b. Site-centered fingerprints pooled to crystal (mean/sum/max)

4. **Graph-based (structure-aware)**
   a. Crystal graphs with atom/bond features (CGCNN/MEGNet/MAT inputs)
   b. Message-passing–derived latent embeddings

Descriptors for Inorganic Materials

1. **Electronic/phononic** (often from DFT)
   a. Formation energy, band gap/DOS features, effective masses, dielectric tensor
   b. Phonon frequencies/DOS, Debye temp, Grüneisen, thermal conductivity surrogates

2. **Mechanical/thermoelastic**
   a. Elastic tensor (Cij), bulk/shear moduli, Poisson ratio, hardness proxies
   b. Thermal expansion, heat capacity

3. **Surfaces/defects/interfaces**
   a. Surface energies, work function, adsorption site descriptors
   b. Defect formation energies, charge transition levels, dopability metrics
   c. Interface misfit/strain, heterostructure registry

Example: MagPie

Magpie key	What it is (per element)	Typical units / notes
Number	Atomic number Z	–
MendeleevNumber	Position in Mendeleev ordering (similarity proxy)	–
AtomicWeight	Atomic mass	amu
MeltingT	Melting temperature	K
Column	Periodic-table group	1–18
Row	Period (row)	1–7
CovalentRadius	Covalent bond radius	Å
Electronegativity	Pauling electronegativity	dimensionless
NsValence	Valence-shell $s$electrons	0–2
NpValence	Valence-shell $p$electrons	0–6
NdValence	Valence-shell $d$electrons	0–10
NfValence	Valence-shell $f$electrons	0–14
NValence	Total valence electrons $= n_s + n_p + n_d + n_f$	0–32
NsUnfilled	Unfilled $s$capacity $= 2 - n_s$)clipped ≥ 0)	0–2
NpUnfilled	Unfilled $p$capacity $= 6 - n_p$	0–6
NdUnfilled	Unfilled $d$capacity $= 10 - n_d$	0–10
NfUnfilled	Unfilled $f$capacity $= 14 - n_f$	0–14
NUnfilled	Total unfilled valence $= (2 - n_s) + (6 - n_p) + (10 - n_d) + (14 - n_f)$	0–32
GSbandgap	Elemental ground-state band gap	eV (≈0 for metals)
GSmagmom	Elemental ground-state magnetic moment	μB/atom
GSvolume_pa	Ground-state volume per atom	Å³/atom
SpaceGroupNumber	Space group of elemental ground state	1–230

npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

A general-purpose machine learning framework for predicting properties of inorganic materials

Logan Ward[1], Ankit Agrawal[2], Alok Choudhary[2] and Christopher Wolverton[1]

A very active area of materials research is to devise methods that use machine learning to automatically extract predictive models from existing materials data. While prior examples have demonstrated successful models for some applications, many more applications exist where machine learning can make a strong impact. To enable faster development of machine-learning-based models for such applications, we have created a framework capable of being applied to a broad range of materials data. Our method works by using a chemically diverse list of attributes, which we demonstrate are suitable for describing a wide variety of properties, and a novel method for partitioning the data set into groups of similar materials to boost the predictive accuracy. In this manuscript, we demonstrate how this new method can be used to predict diverse properties of crystalline and amorphous materials, such as band gap energy and glass-forming ability.

npj Computational Materials (2016) **2,** 16028; doi:10.1038/npjcompumats.2016.28; published online 26 August 2016

From descriptors to properties

Hume–Rothery (alloys: solid solutions & e⁄a phases) s**olid-solution rules:**

- **Atomic size mismatch ≤ ~15%** → high solubility
- **Same crystal structure** (e.g., fcc with fcc)
- **Similar electronegativity** (avoids compound formation)
- **Same valence** mixes more readily

•**Electron phases:** β/γ/ε stabilize near characteristic **e⁄a** ratios (~1.5–1.75).

•**Use:** Alloy design, brass/bronze families. **Caveat:** Many exceptions with directional/ionic bonding.

Inoue Rules (bulk metallic glasses, 1990s): Enhance glass-forming ability (GFA) in metallic systems.

- **Multi-component** (≥3 principal elements)
- **Large atomic size mismatch** (≳ **12%**) among main constituents
- **Negative heats of mixing** between pairs → deep eutectics, sluggish crystallization

•**Use:** Design of BMGs/HEA-like glasses. **Caveat:** Processing path (cooling rate) is critical.

Pauling's Five Rules (ionic crystals, 1929)

1. **Coordination (radius-ratio)** sets coordination number.
2. **Electrostatic valence:** bond strengths to each anion **sum to its charge**.
3. **Polyhedra sharing:** corner-sharing favored; **edge/face sharing destabilizes**, esp. for small/high-valent cations.
4. **Avoid sharing** of polyhedra around **small, highly charged cations**.
5. **Parsimony:** crystals tend to have **few distinct coordination environments**.

•**Use:** First-pass structure rationalization for salts/oxides.

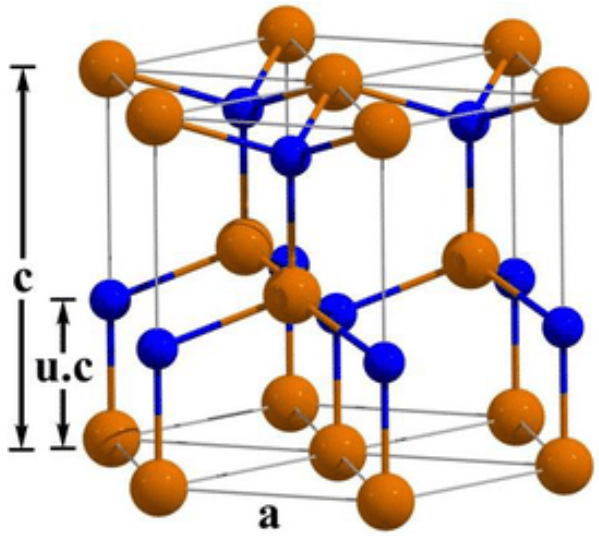
•**Caveat:** Covalency, polarization, lone pairs can violate.

Wait, there is more….

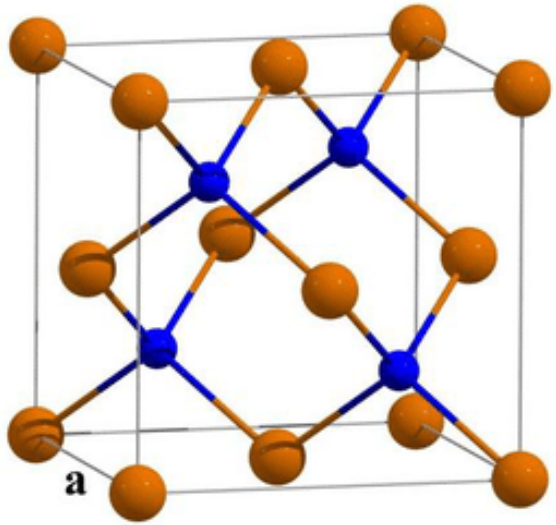
Rule	Domain	One-line heuristic
Zachariasen rules	Oxide glasses	Glass networks from **low-coordination formers**; **corner-sharing** units; O atoms mostly **bridge two cations**; avoid O–O bonds.
Goldschmidt tolerance (± octahedral factor)	Perovskites (ABO_3)	Stable when **t ≈ 0.8–1.05** and **μ = r_B/r_O** in allowed range; predicts tilts/formability.
Mooser–Pearson / 8–N	Main-group semiconductors	**Structure/ionicity** trend with **valence electron count**; coordination often follows **8–N**.
Goodenough–Kanamori–Anderson	Oxide magnetism (superexchange)	**Coupling sign/magnitude** set by **orbital occupancy** and **M–O–M angle** (≈180° → AFM; ~90° can favor FM).
HEA heuristics (Ω, δ, VEC)	High-entropy alloys	Single-phase solid solution when **size mismatch δ small**, **Ω = T_mΔS_mix/
Vegard's law	Alloy lattice parameters	**Lattice constant ~ linear** in composition (first-order mix).
Sabatier/volcano	Heterogeneous catalysis	Best catalysts bind intermediates **neither too weakly nor too strongly** → **volcano** vs binding energy.
Confusion principle	Glassy/complex alloys	**Many competing crystalline states** (similar energies) **frustrate crystallization** → glass/solid solution.
Miedema model heuristics	Alloy mixing enthalpy	Predict **ΔH_mix** from **work function & electron density** parameters → sign and magnitude trends.

Binary Octet Compounds

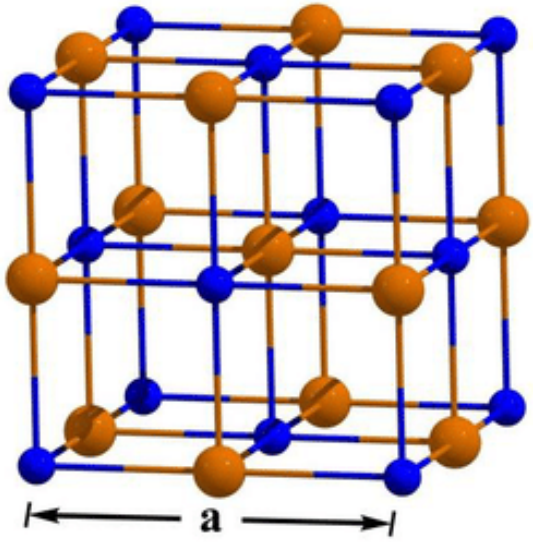
- NaCl, LiI, BeO, AlN, ….
- Can exist in zincblende (ZB), wurtzite (WZ), rocksalt (RS), cesium chloride (CsCl), and diamond cubic (DC) crystal structures



(a) wurtzite



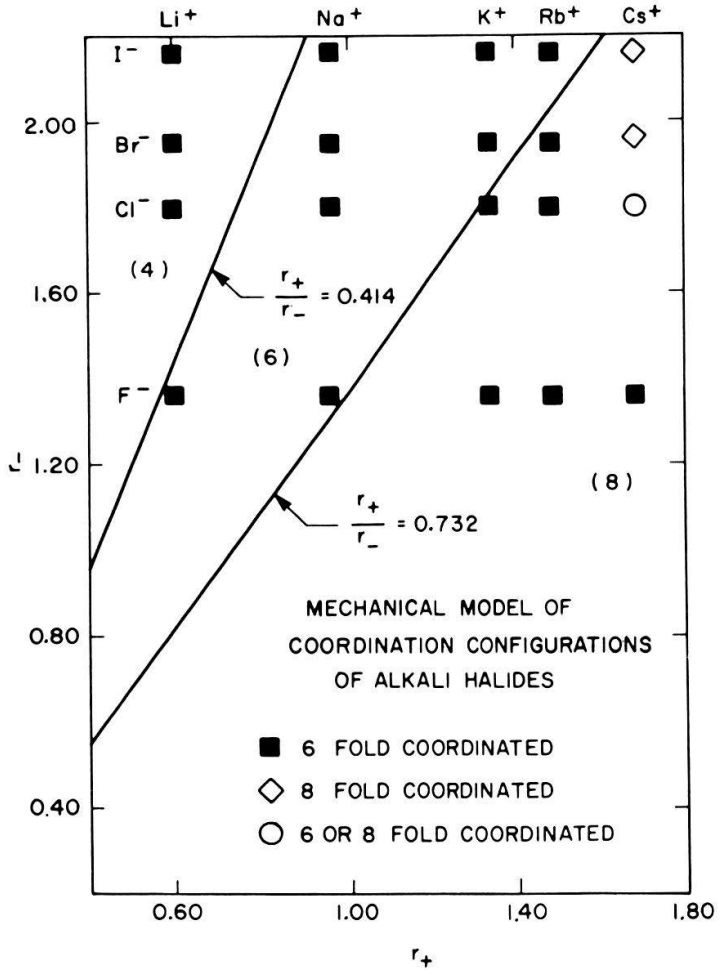
(b) zinc-blende



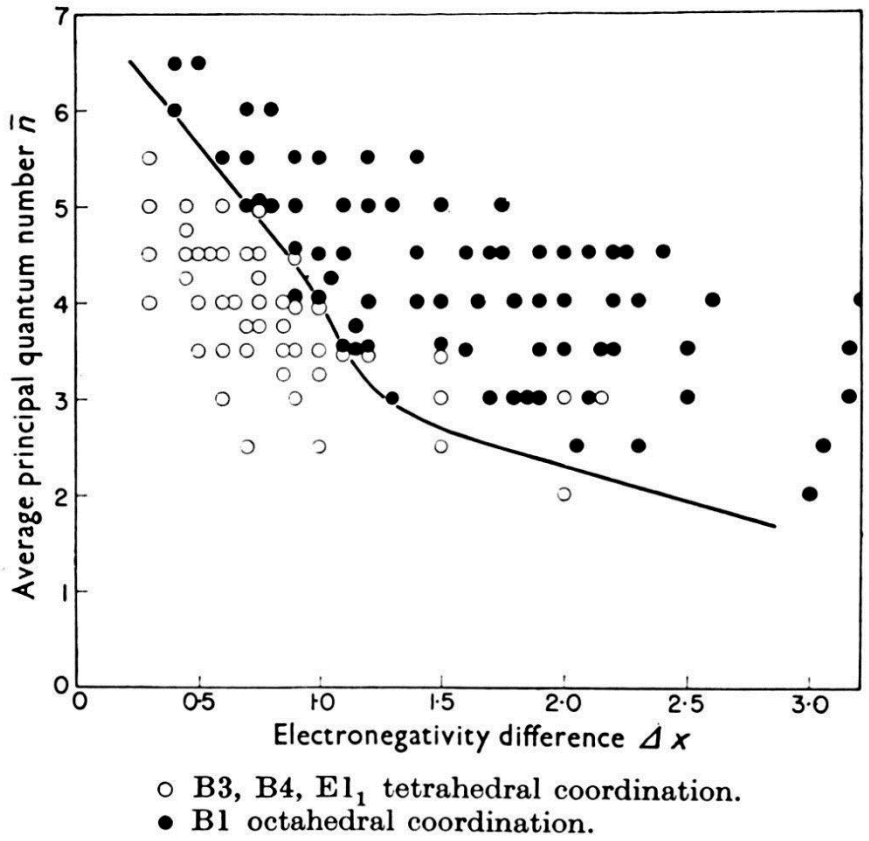
(c) rock-salt

T. Wonglakhon and D. Zahn, Interaction Potentials for modelling GaN precipitation and solid state polymorphism, Journal of Physics Condensed Matter 32(20), DOI:10.1088/1361-648X/ab6cbe

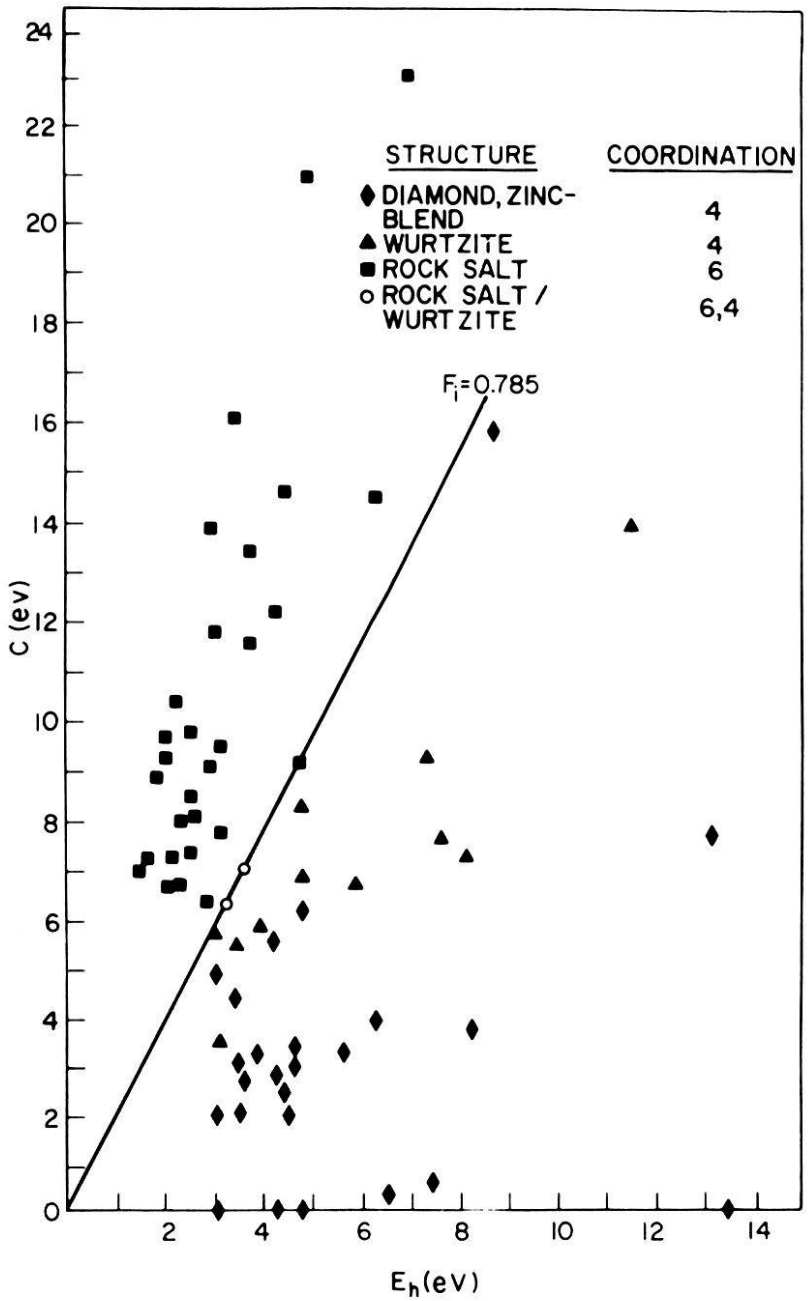
Can we predict the structure from composition?



Mooser-Pearson plots, 1959



J. C. Phillips, Structure and Properties: Mooser-Pearson plots, Helvetica Physica Acta, Vol. 58 (1985)



This average energy gap E_g was separated into covalent and ionic components, E_h and C respectively, by a Hückel relation $E_g^2 = E_h^2 + C^2$. One could then determine E_h and C separately by scaling the former with the bond length d and obtain E_g and C from ϵ. In this model the transformation from tetrahedral to octahedral coordination depends on the fraction of ionic character in the chemical bond given by $f_i = C^2/E_g^2$.

The Phillips-Van Vechten plot for AB valence compounds utilizing 'symmetric' energy-gap coordinates Eh and C. The use of quantum-mechanically defined coordinates, together with the restriction to valence compounds and exclusion of transition-metal compounds, leads to an exact separation with a straight line corresponding to constant critical ionicity.

J. C. Phillips, Structure and Properties: Mooser-Pearson plots, Helvetica Physica Acta, Vol. 58 (1985)

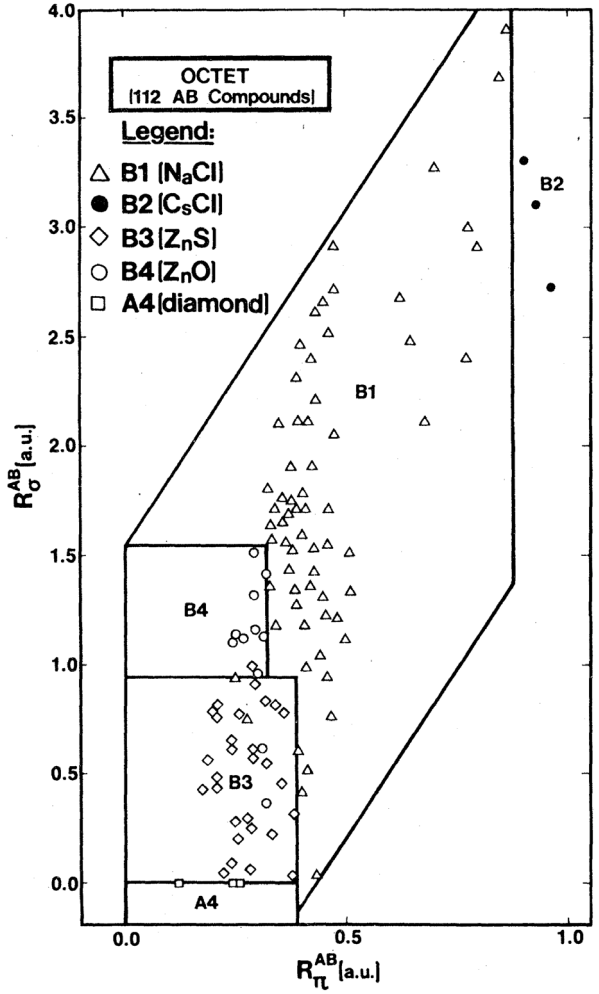


FIG. 19. Structural separation plot for the 112 binary octet compounds $A^N B^{(8-N)}$, obtained with the density-functional orbital radii, with

$$R_\sigma^{AB} = |\,(r_p^A + r_s^A) - (r_p^B + r_s^B)\,|,$$

$$R_\pi^{AB} = |\,r_p^A - r_s^A\,| + |\,r_p^B - r_s^B\,|.$$

Villars diagrams

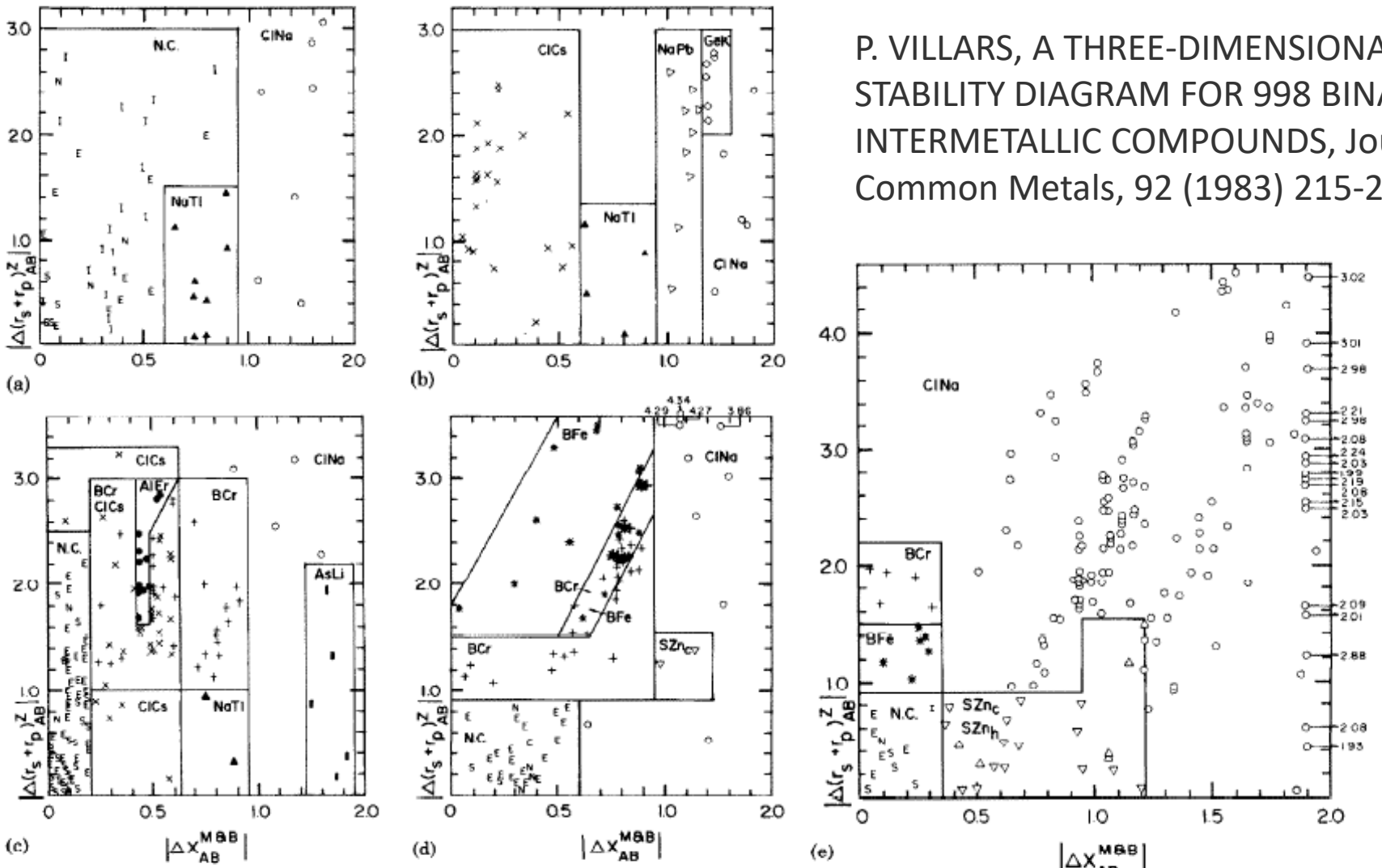
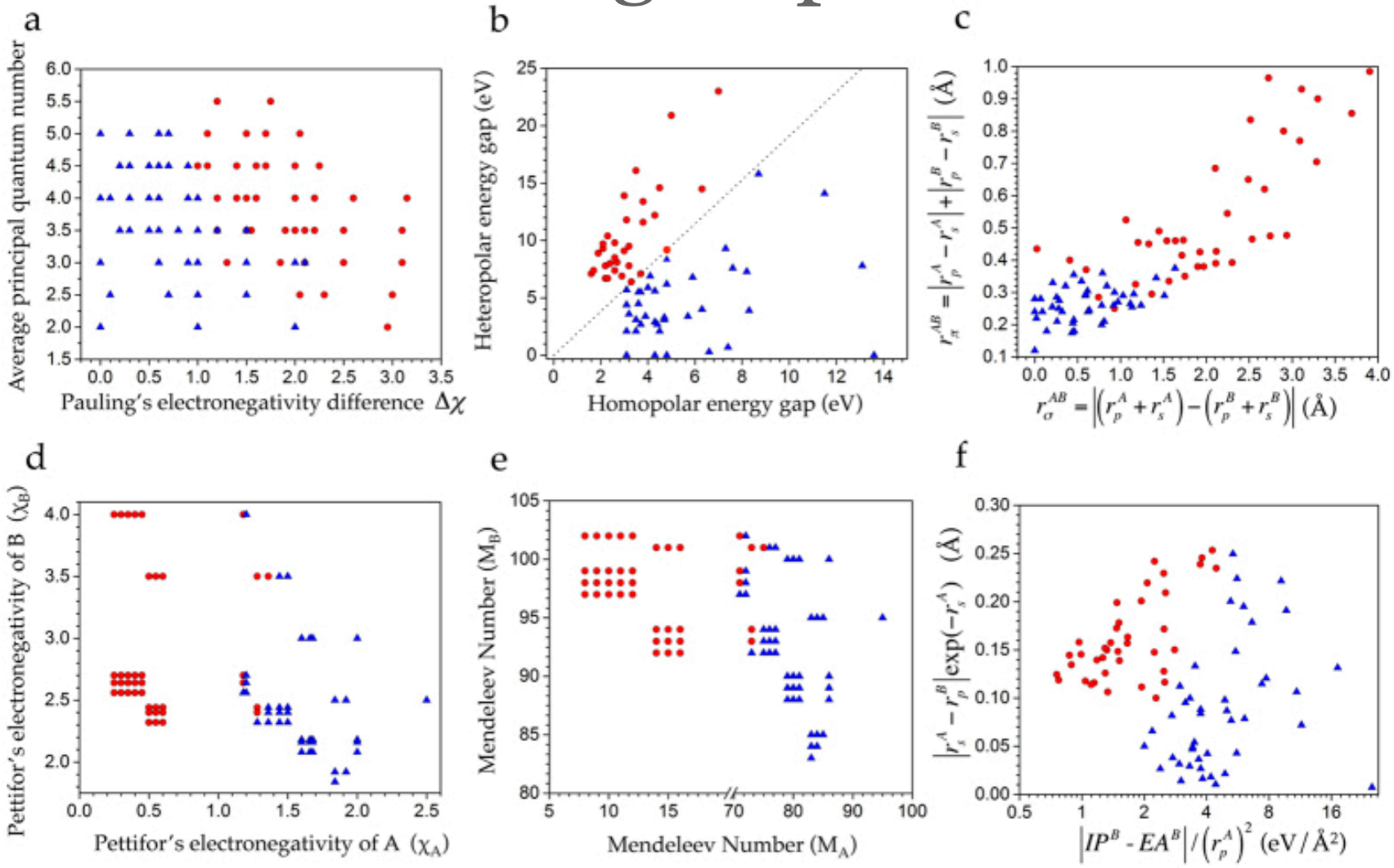


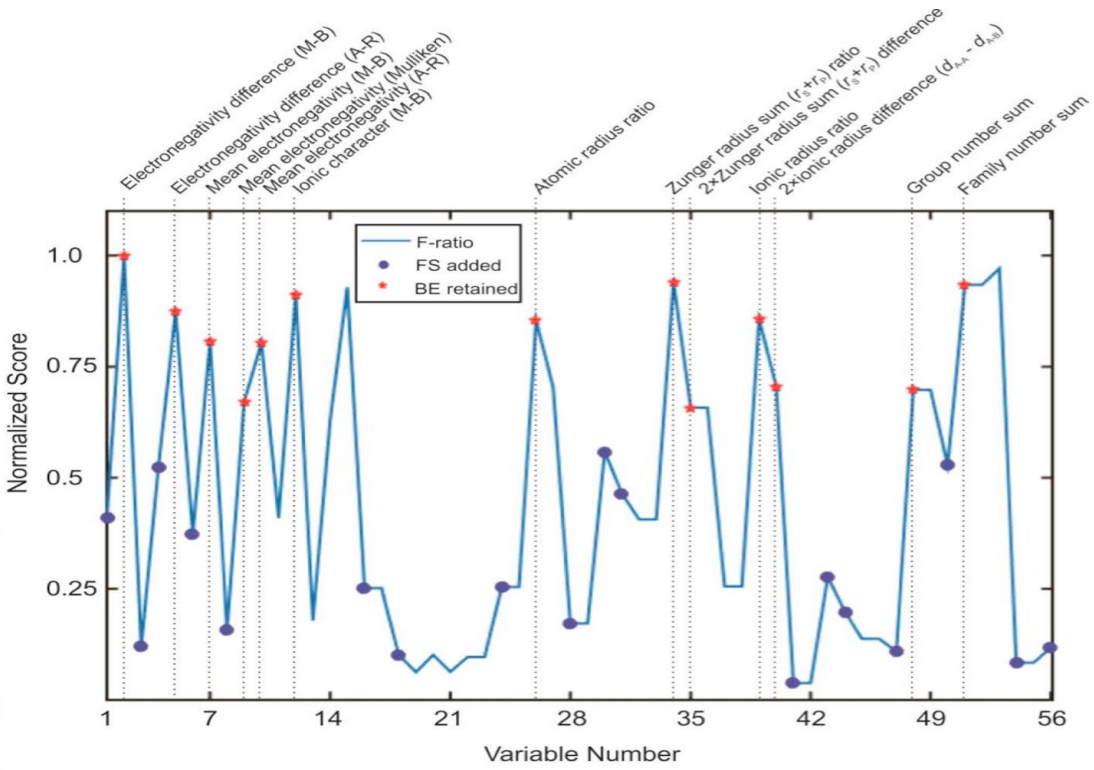
Fig. 3 *(continued).*

P. VILLARS, A THREE-DIMENSIONAL STRUCTURAL STABILITY DIAGRAM FOR 998 BINARY AB INTERMETALLIC COMPOUNDS, Journal of the Less-Common Metals, 92 (1983) 215-238 215

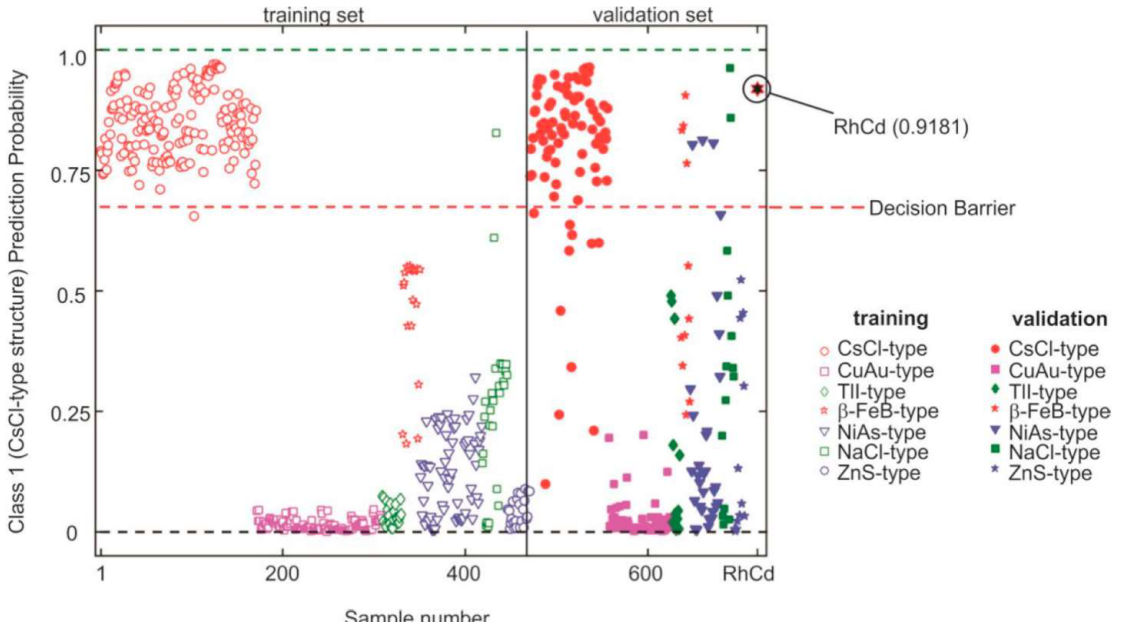
Can machine learning help?



G. Pilania, J. E. Gubernatis, and T. Lookman, Classification of octet AB-type binary compounds using dynamical charges: A materials informatics perspective, Sci Rep. 2015; 5: 17504.



1. ●Electronegativity difference (Pauling scale)	13. Ionic character (Gordy scale)	37. Ionic radius sum (d_{A-B})
2. ★Electronegativity difference (Martynov-Batsanov scale)	14. Ionic character (Mulliken scale)	38. Mean ionic radius
3. ●Electronegativity difference (Gordy scale)	15. Ionic character (Allred-Rochow scale)	39. ★Ionic radius ratio
4. ●Electronegativity difference (Mulliken scale)	16. ●Sum of valence electrons	40. ★2×ionic radius difference (d_{A-A} - d_{A-B})
5. ★Electronegativity difference (Allred-Rochow scale)	17. Mean number of electrons	41. ●Crystal radius sum (d_{A-B})
6. ●Mean electronegativity (Pauling scale)	18. ●Atomic number sum	42. Mean crystal radius
7. ★Mean electronegativity (Martynov-Batsanov scale)	19. Atomic number difference	43. ●Crystal radius ratio
8. ●Mean electronegativity (Gordy scale)	20. Mean atomic number	44. ●2×crystal radius difference (d_{A-A} - d_{A-B})
9. ★Mean electronegativity (Mulliken scale)	21. Atomic weight difference	45. Period number sum
10. ★Mean electronegativity (Allred-Rochow scale)	22. Mean atomic weight	46. Mean period number
11. Ionic character (Pauling scale)	23. Atomic weight sum	47. ●Period number difference
12. ★Ionic character (Martynov-Batsanov scale)	24. ●Atomic radius sum (d_{A-B})	48. ★Group number sum
	25. Mean atomic radius	49. Mean group number
	26. ★Atomic radius ratio	50. ●Group number difference
	27. 2×atomic radius difference (d_{A-A} - d_{A-B})	51. ★Family number sum
	28. ●Covalent radius sum (d_{A-B})	52. Mean Family number
	29. Mean covalent radius	53. Family number difference
	30. ●Covalent radius ratio	54. ●Quantum number (l) sum
	31. ●2×covalent radius difference (d_{A-A} - d_{A-B})	55. Mean quantum number (l) mean
	32. Zunger radius sum (r_S+r_P) sum	56. ●Quantum number (l) difference
	33. Mean Zunger radius sum (r_S+r_P)	
	34. ★Zunger radius sum (r_S+r_P) ratio	
	35. ★2×Zunger radius sum (r_S+r_P) difference	
	36. Zunger radius sum (r_S+r_P) difference	



A. O. Oliynyk, L.A. Adutwum, J.J. Harynuk, and A. Mar, Classifying Crystal Structures of Binary Compounds AB through Cluster Resolution Feature Selection and Support Vector Machine Analysis, Chem. Mater. 2016, 28, 18, 6672–6681 (2016)

Feature engineering with machine learning

For instance, the starting point $\mathbf{\Phi}_0$ may comprise readily available and relevant properties, such as atomic radii, ionization energies, valences, bond distances, and so on. The operators set is defined as

$$\hat{H}^{(m)} \equiv \{I, +, -, \times, /, \exp, \log, |-|, \sqrt{}, ^{-1}, ^{2}, ^{3}\}[\phi_1, \phi_2],$$

- Start with available physical descriptors
- Create dimensionally-consistent combinations via allowed operations
- Choose the ones that give best classification

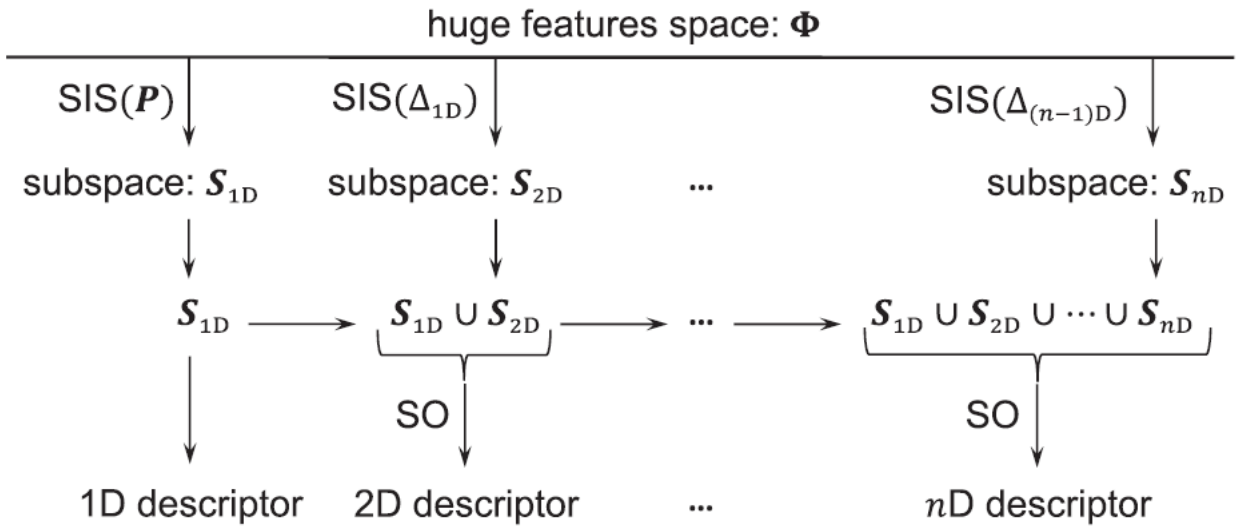


FIG. 1. The method SISSO combines unified subspaces having the largest correlation with residual errors $\mathbf{\Delta}$ (or $\boldsymbol{P}$) generated by sure independence screening (SIS) with sparsifying operator (SO) to further extract the best descriptor.

R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L.M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, PHYSICAL REVIEW MATERIALS 2, 083802 (2018)

Feature engineering with machine learning

TABLE I. Dependence of the metal-insulator classification descriptors on the prototypes of training binary materials.

prototypes	#materials	primary features	descriptor	classification accuracy
NaCl	132	IE_A, IE_B, χ_A, χ_B, r_{covA}, r_{covB}, EA_A, EA_B, v_A, v_B, d_{AB}	$d_1 := \frac{IE_A IE_B (d_{AB} - r_{\text{covA}})}{\exp(\chi_A)\sqrt{r_{\text{covB}}}}$	100%
NaCl, CsCl, ZnS, CaF_2, Cr_3Si	217	IE_A, IE_B, χ_A, χ_B, r_{covA}, r_{covB}, d_{AB}, CN_A, CN_B	$d_1 := \frac{IE_B d_{AB}^2}{\chi_A r_{\text{covA}}^2 \sqrt{CN_B}}$, $d_2 := \frac{IE_A^2 r_{\text{covB}} \log(IE_A) \lvert r_{\text{covA}} - r_{\text{covB}} \rvert}{CN_B}$	100%
NaCl, CsCl, ZnS, CaF_2, Cr_3Si, SiC, TiO_2, ZnO, FeAs, NiAs	260	IE_A, IE_B, χ_A, χ_B, r_{covA}, r_{covB}, d_{AB}, CN_A, CN_B	$d_1 := \frac{d_{AB}/r_{\text{covA}} - \chi_A/\chi_B}{\exp(CN_B/IE_B)}$, $d_2 := \frac{r_{\text{covA}}^3 d_{AB} IE_B}{\lvert \chi_B/\chi_A - \lvert CN_B - CN_A \rvert \rvert}$	99.6%[a]
NaCl, CsCl, ZnS, CaF_2, Cr_3Si, SiC, TiO_2, ZnO, FeAs, NiAs	260	IE_A, IE_B, χ_A, χ_B, x_A, x_B, $V_{\text{cell}}/\sum V_{\text{atom}}$	$d_1 := \frac{V_{\text{cell}}}{\sum V_{\text{atom}}} \frac{\sqrt{\chi_B}}{\chi_A}$, $d_2 := \frac{IE_A IE_B}{\exp(V_{\text{cell}}/\sum V_{\text{atom}})}$	99.6%[a]
NaCl, CsCl, ZnS, CaF_2, Cr_3Si, SiC, TiO_2, ZnO, FeAs, NiAs, Al_2O_3, La_2O_3, Th_3P_4, ReO_3, ThH_2	299	IE_A, IE_B, χ_A, χ_B, x_A, x_B, $V_{\text{cell}}/\sum V_{\text{atom}}$	$d_1 := \frac{x_B}{\sum V_{\text{atom}}/V_{\text{cell}}} \frac{IE_B \sqrt{\chi_B}}{\chi_A}$, $d_2 := \chi_A^2 \lvert \lvert 1 - 2x_A \rvert - x_A^2 \frac{\chi_B}{\chi_A} \rvert$	99.0%[b]

[a]One entry misclassified: YP compound in NaCl prototype.

[b]Three entry misclassified: YP compound in NaCl prototype; Th_3As_4 and La_3Te_4 compounds in Th_3P_4 prototype.

R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L.M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, PHYSICAL REVIEW MATERIALS 2, 083802 (2018)

Feature engineering with machine learning

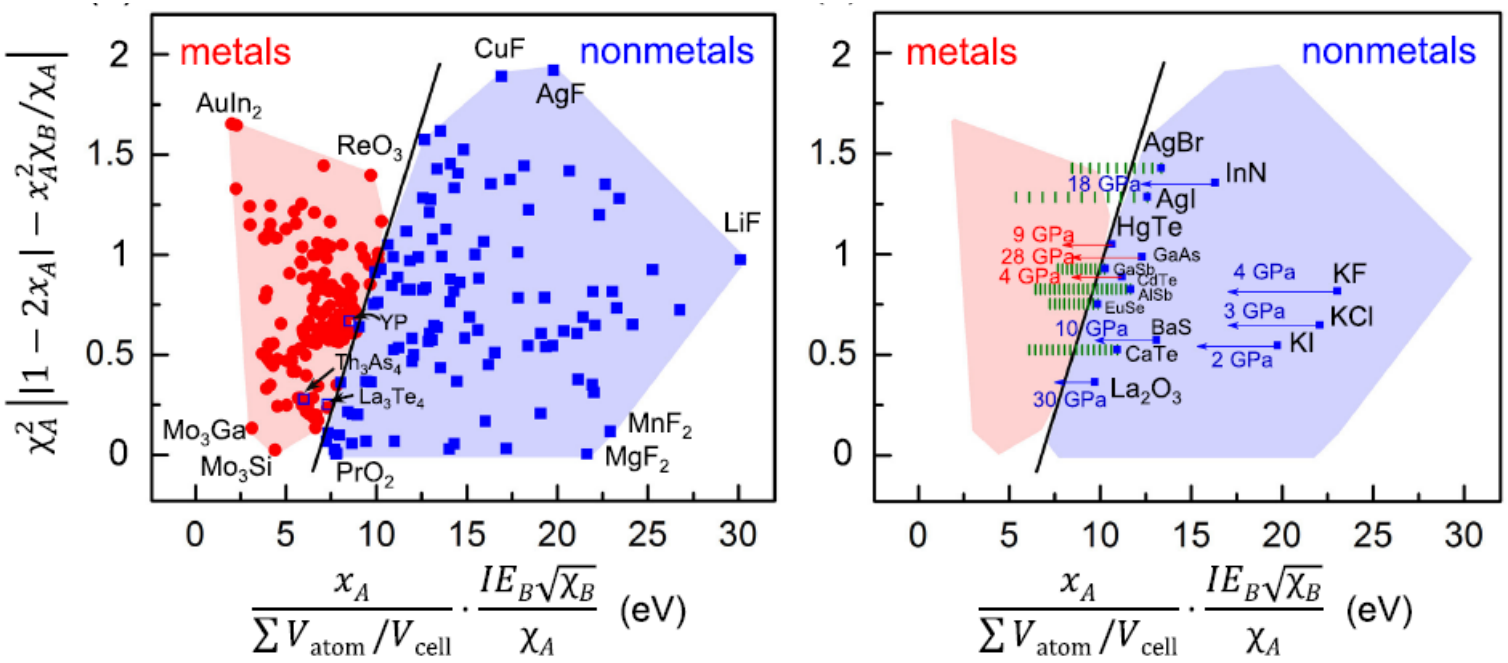


FIG. 4. SISSO for classification. (a) An almost perfect classification (99%) of metal/nonmetal for 299 materials. Symbols: χ, Pauling electronegativity; IE, ionization energy; x, atomic composition; $\sum V_{\text{atom}}/V_{\text{cell}}$, packing fraction. Red circles, blue squares, and open blue squares represent metals, nonmetals, and the three erroneously characterized nonmetals, respectively. (c) Reproduction of pressure-induced insulatormetals transitions (red arrows), of materials that remain insulators upon compression (blue arrows), and computational predictions at step of 1 GPa (green bars).

R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L.M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, PHYSICAL REVIEW MATERIALS 2, 083802 (2018)

Inorganic Materials

1. Synthesis & "synthesizability"

- Route/precursor recommendation; reaction condition prediction
- Likelihood of successful synthesis; lab yield/phase purity forecasting

2. Microstructure–property links

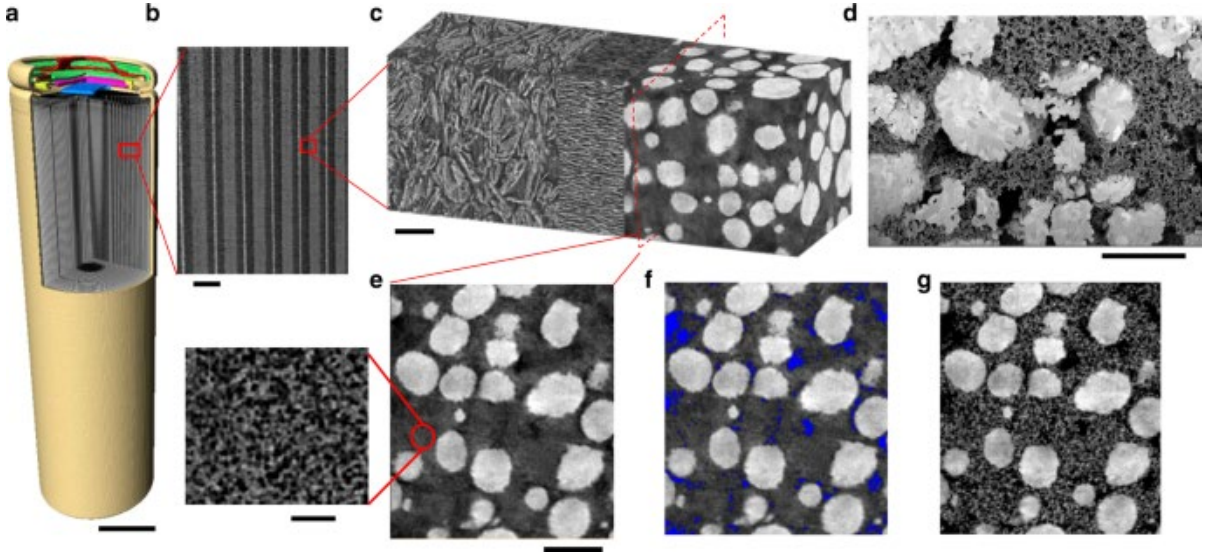
- Predict properties from grain/texture/porosity descriptors
- Aging/degradation/lifetime under stress, environment, cycling

3. Experimental interpretation

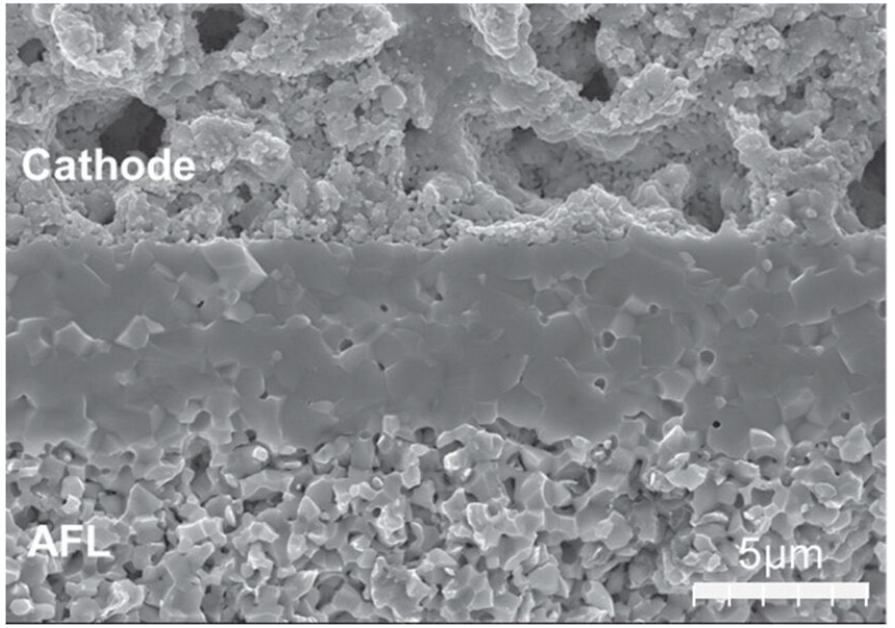
- Automated **XRD/XAS/EELS** phase ID and quantification
- Real-time decision support in autonomous labs

Microstructures

https://manyeats.com/damascus-steel-and-its-modern-attempts/

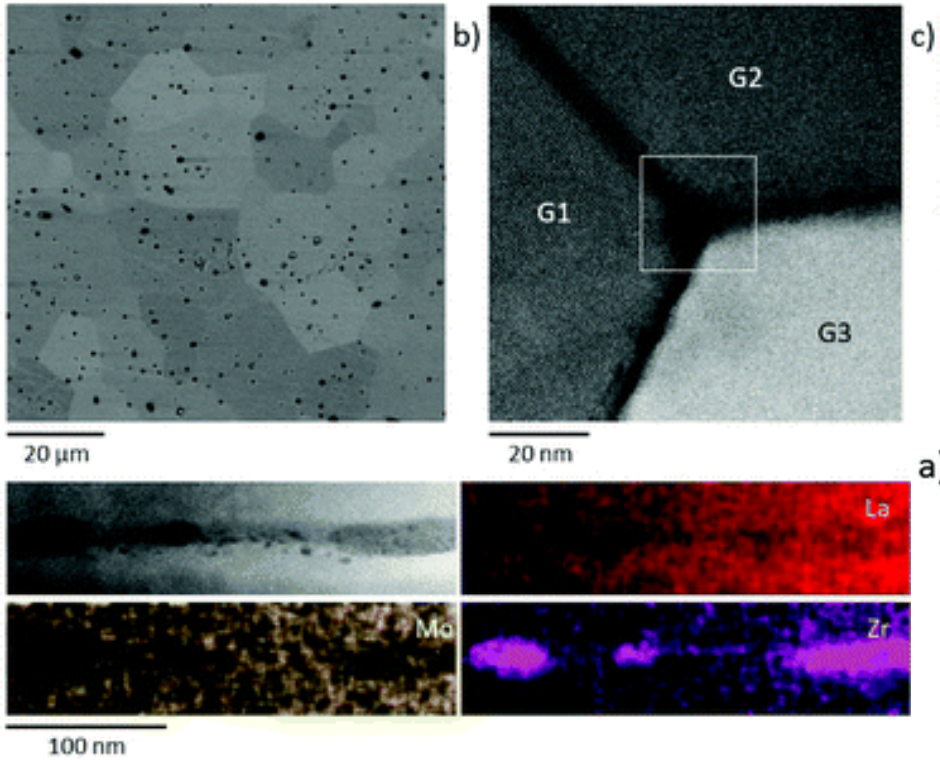


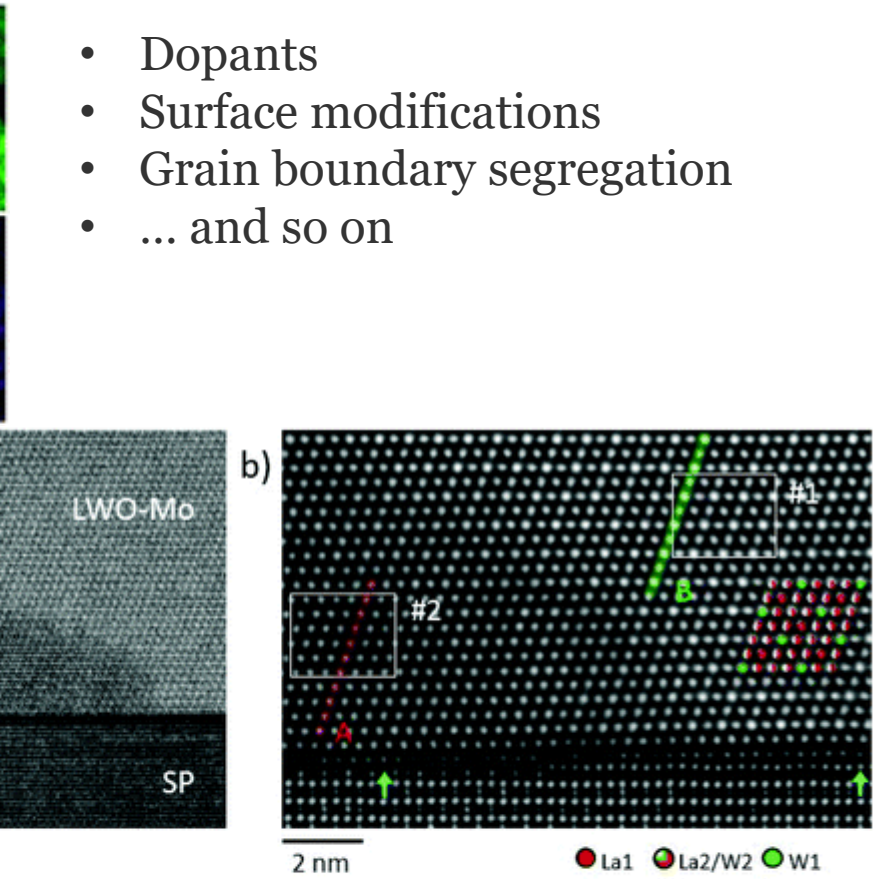
https://www.nature.com/articles/s41467-020-15811-x



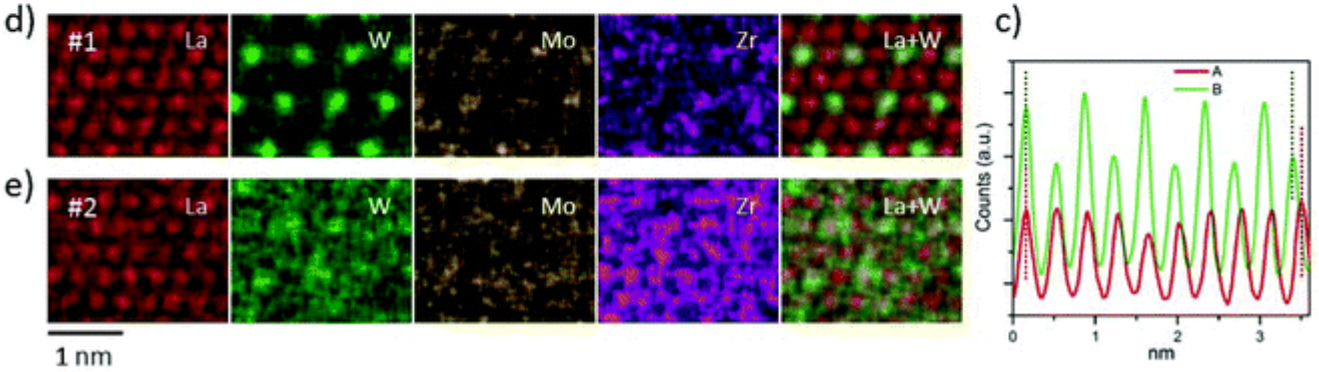
https://www.researchgate.net/figure/Microstructure-of-the-electrolyte-electrode-interface-region-in-the-fuel-cell-with-

Impurities Matter





- Dopants
- Surface modifications
- Grain boundary segregation
- … and so on

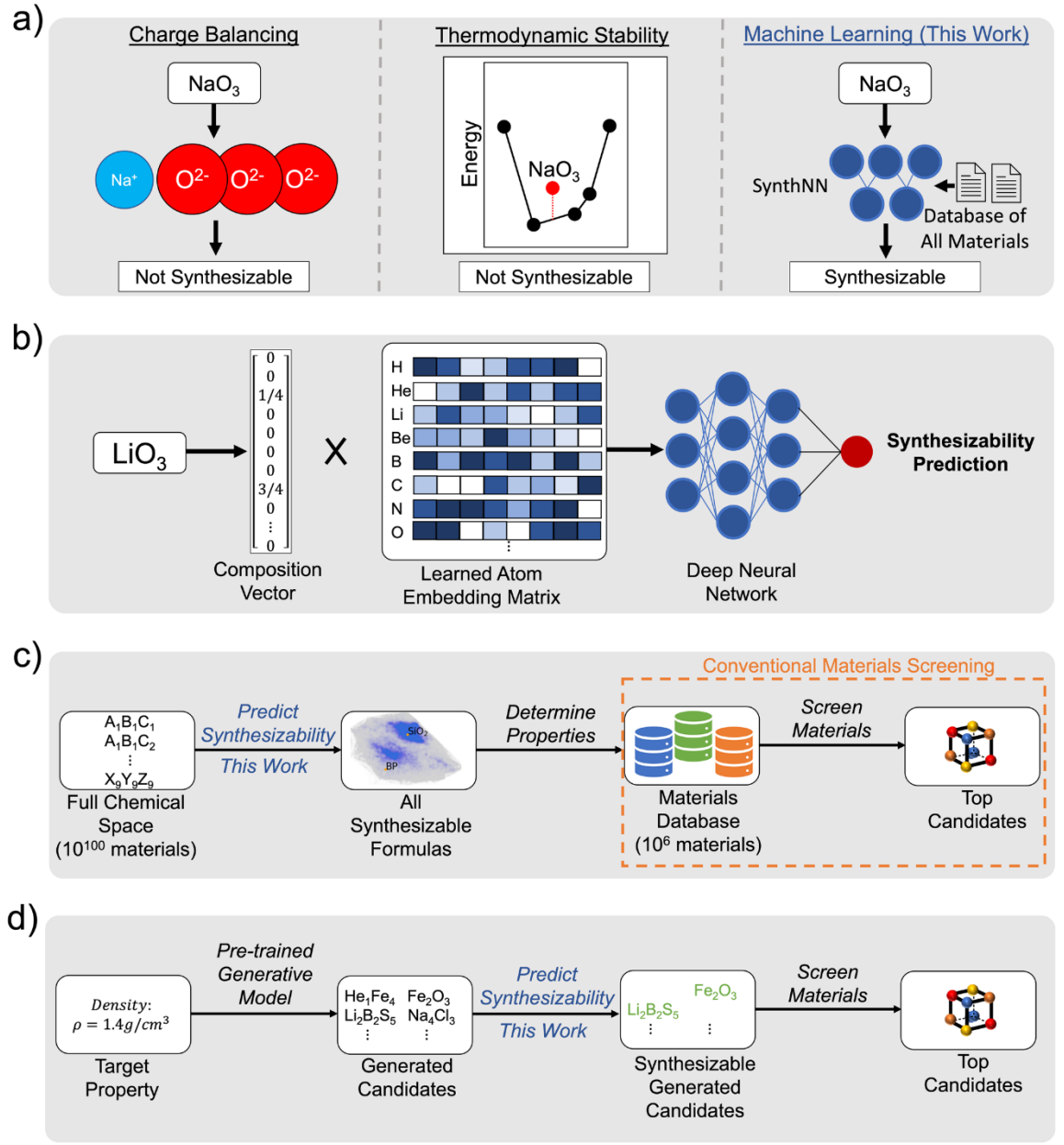


https://pubs.rsc.org/en/content/articlelanding/2020/nr/d0nr04396f

1. **Synthesis: initial reagents and processing trajectory**

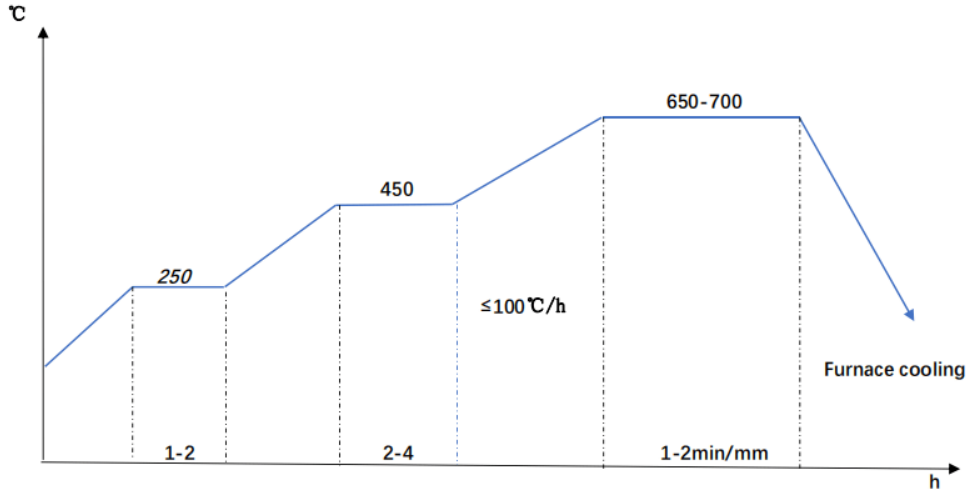
2. **Domain specific and expressed in terms of operations in real laboratories**

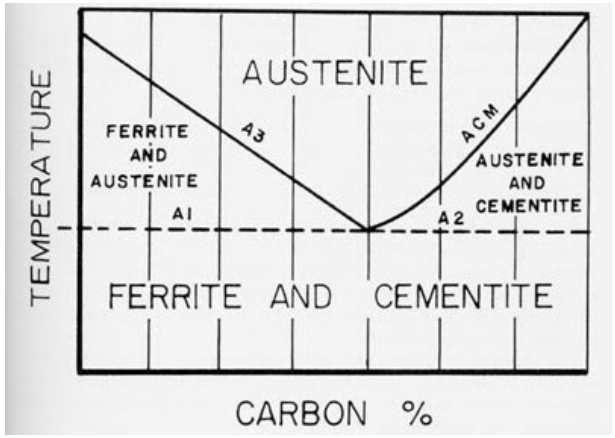
3. **Where would new operations come from?**

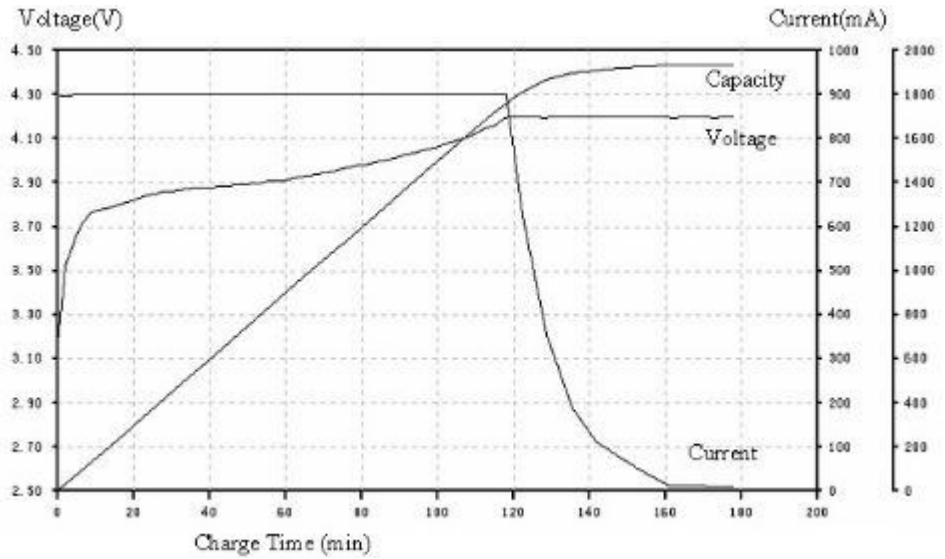


https://www.nature.com/articles/s41524-023-01114-4

Making materials: process trajectories







- Making steel: complicated and took a lot of time optimize
- Charging battery: obvious economic impact
- Manufacturing: Annealing hybrid perovskite thin films
- Poling ferroelectric

How do we optimize trajectories if we have (a) only limited or no mechanistic information, (b) our experimental budgets are limited, but (c) we have some access to domain expertise?

Why dimensionality is a problem?

- Suppose that we have data for 1000 students' performance (discretized scores of 0; 25; 50; 75; 100)% in 2 courses c1 and c2. Then in total there are 5 x 5 = 25 different grade combinations.

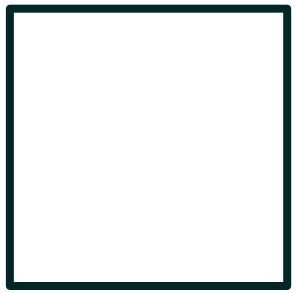
- If the 1000 students are randomly distributed among each grade combination, then on average there are 40 students with each possible grade combination, which is a good enough sample to draw conclusions such as if, for a student, grade(c1) 50 and grade(c2) 75, then that student is likely to be a Math major.

- Now suppose there are 4 courses, then the number of possible grades combination is 54 = 625, and an average number of students per combination is 1:6. For 10 courses, this number reduces to 0:0001024. This means that almost all possible combinations are never observed.

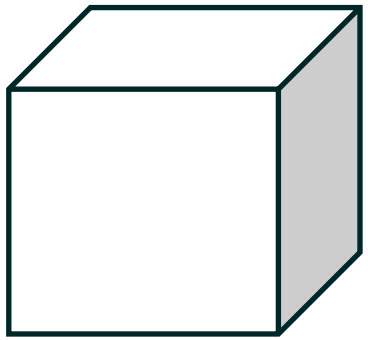
From "Curse of Dimensionality", Lecture Notes for Big Data Analytics, Faizad Ullah, February 2019

Why dimensionality is a problem?

- Suppose *n* points in X are chosen uniformly at random from $[0; 1]^m$ (m-cube). For the query point q grow a hypercube around *q* to contain *f* fraction of points ($k = f\ n$) in X. This cube (the search space for q) grows very large (covering almost the whole input space) in large dimension.

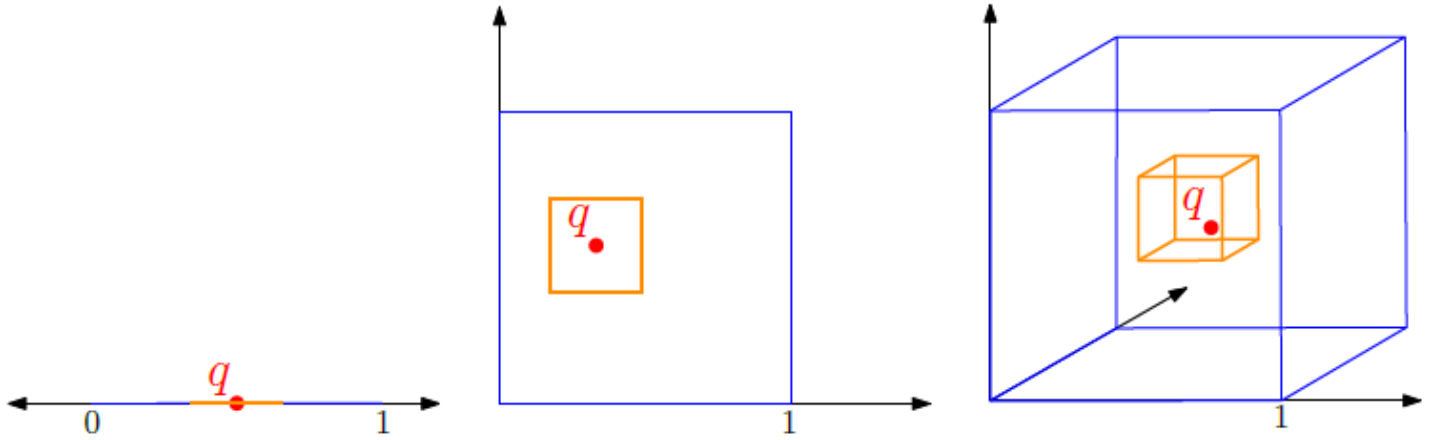


- The expected length of the edge of the search cube $E_m(f) = f^{1/m}$, i.e. in 10d to get 10% points around q need cube with edge length 0.8 (which is 80% of the whole cube, the input space). Similarly, to get only 1% points one needs to extend the search cube by 0:63 units along each dimension



From “Curse of Dimensionality”, Lecture Notes for Big Data Analytics, Faizad Ullah, February 2019

Why dimensionality is a problem?

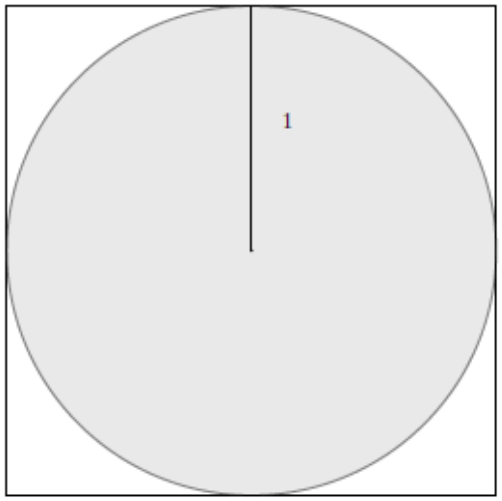


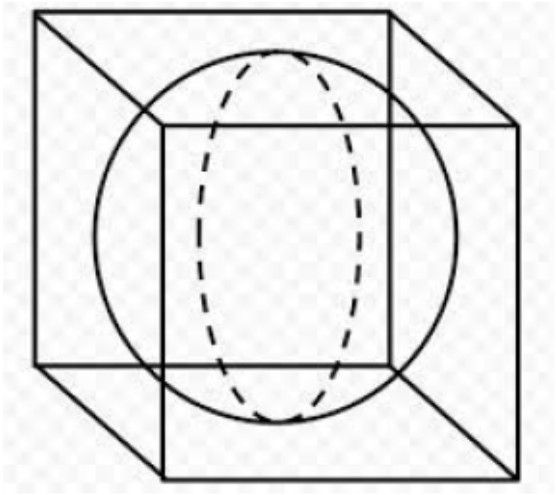
Suppose we have 5000 points:

- In 1d we have to explore 0.001 on average to capture 5 NN
- In 2d, on average we must explore 0:031 units along both dimensions to get 5 nearest neighbors points (about 3% of the whole cube).
- In 3d, on average we must go 10% of the total (unit) length in each of the 3 dimensions
- In 4d, we must explore 17:7% of unit length
- In 10d, we must go 50.1% of unit length along each dimension

From “Curse of Dimensionality”, Lecture Notes for Big Data Analytics, Faizad Ullah, February 2019

Why dimensionality is a problem?





dim m	volume of m-ball	volume of m-cube	ratio
2	π	2^2	~ 0.785
3	$4/3\pi$	2^3	~ 0.523
4	$\pi^2/2$	2^4	~ 0.308
6	$\pi^3/6$	2^6	~ 0.080
m	$\frac{\pi^{m/2}}{m/2!}$	2^m	$\to 0$

From “Curse of Dimensionality”, Lecture Notes for Big Data Analytics, Faizad Ullah, February 2019

Why dimensionality is a problem?

However if a dataset exhibit this phenomenon that the issue has be overcome by getting a larger training set (exponential in m). One way to look at this is as follows.

To cover $[-1,1]^m$ with $B_{m,1}$'s, the number of balls n must be

$$n \geq \frac{2^m}{V_m(1)} = \frac{2^m}{\pi^{m/2}/m/2!} = \frac{m/2!\, 2^m}{\pi^{m/2}} \overset{m\to\infty}{\sim} \sqrt{m\pi}\left(\frac{m2^{m/2}}{2\pi e}\right)^{m/2}$$

For $m = 16$ (a very small number) this n is substantially larger than 2^{58}

- In higher dimensions all the volume is in `corners'
- Points in high dimensional spaces are isolated (empty surrounding)
- The probability that a randomly generated point is within r radius of q approaches 0 as dimensionality increases
- The probability of a close nearest neighbor in a data set is very small

From “Curse of Dimensionality”, Lecture Notes for Big Data Analytics, Faizad Ullah, February 2019

From descriptors to properties

Rule (origin)	Domain / scope	Core heuristic (summary)	Typical use	Caveats
Hume–Rothery (1920s–50s)	Substitutional solid solutions & electron phases in alloys	(i) **Size** mismatch ≲ **15%** for unlimited solubility; (ii) **Same crystal structure** favors solubility; (iii) **Similar electronegativity** to avoid compound formation; (iv) **Valence**: metals of same valence mix better. *Electron-concentration phases*: β, γ, ε phases stabilize near characteristic **e⁄a** ratios (~1.5, ~1.62, ~1.75).	Predicting solubility, phase formation, and brass-type phases in alloys.	Empirical; many exceptions with directional/ionic bonding, strong ordering, or complex intermetallics.
Pettifor chemical scale & structure maps (1980s–90s)	Binary/ternary intermetallics & ceramic structure types	Order elements by **Pettifor/Mendeleev number**; plot binaries on maps—**structure types** (e.g., B1/B2/B3, NiAs) cluster in regions → interpolate/extrapolate likely structures.	Rapid prototype/structure selection; screening composition–structure relationships.	Map depends on chosen scale; boundaries are fuzzy; kinetics and defects not captured.
Inoue rules (1990s) for **bulk metallic glass (BMG) formation**	Glass-forming ability in metallic systems	(i) **Multi-component** (≥3 elements); (ii) Large **atomic size mismatch** (≳12% among main constituents); (iii) **Negative heats of mixing** between pairs → deep eutectics, sluggish crystallization.	Designing BMGs and HEA-like glassy alloys.	Composition windows narrow; processing (cooling rate) critical; not sufficient alone.
Zachariasen rules (1932)	**Oxide glasses**	Network formers (Si, B, P) with **low coordination**; **corner-sharing** polyhedra; each O bridges **two cations**; **no O–O bonds**.	Identifying glass formers / modifiers in oxides.	Qualitative; many modified/complex glasses deviate.
Pauling's five rules (1929)	**Ionic crystals**	(1) **Radius-ratio**→coordination; (2) **Electrostatic valence** sum at each anion; (3–4) **Edge/face sharing** of cation polyhedra destabilizes (esp. small/high-valent cations); (5) **Parsimony**: few distinct sites.	First-pass structure rationalization for salts/oxides.	Oversimplifies covalency, polarization, and lone-pair effects.
Goldschmidt tolerance (1926) (+ **octahedral factor**)	**Perovskites** ABO_3	**$t = (r_A+r_O)/\sqrt{2}(r_B+r_O) \approx 0.8–1.05$** stable; **$\mu = r_B/r_O$** in ~0.41–0.73.	Screening A/B cations for perovskite formability; tilt trends.	Ionic-radius choice matters; breaks for hybrid/strongly covalent or distorted systems.
	Zincblende, wurtzite, rocksalt	Structure & bonding trend with	Quick structure/bonding guess in	Semi-empirical; d/f participation and

From descriptors to properties

Rule (origin)	Domain / scope	Core heuristic (summary)	Typical use	Caveats
Hume–Rothery (1920s–50s)	Substitutional solid solutions & electron phases in alloys	**Size** mismatch ≲ **15%** for unlimited solubility; (ii) **Same crystal structure** favors solubility; (iii) **Similar electronegativity** to avoid compound formation; (iv) **Valence**: metals of same valence mix better. *Electron-concentration phases*: β, γ, ε phases stabilize near characteristic **e⁄a** ratios (~1.5, ~1.62, ~1.75).	Predicting solubility, phase formation, and brass-type phases in alloys.	Empirical; many exceptions with directional/ionic bonding, strong ordering, or complex intermetallics.
Pettifor chemical scale & structure maps (1980s–90s)	Binary/ternary intermetallics & ceramic structure types	Order elements by **Pettifor/Mendeleev number**; plot binaries on maps—**structure types** (e.g., B1/B2/B3, NiAs) cluster in regions → interpolate/extrapolate likely structures	Rapid prototype/structure selection; screening composition–structure relationships.	Map depends on chosen scale; boundaries are fuzzy; kinetics and defects not captured.
Inoue rules (1990s) for **bulk metallic glass (BMG) formation**	Glass-forming ability in metallic systems	(i) **Multi-component** (≥3 elements); (ii) Large **atomic size mismatch** (≳12% among main constituents); (iii) **Negative heats of mixing** between pairs → deep eutectics, sluggish crystallization.	Designing BMGs and HEA-like glassy alloys.	Composition windows narrow; processing (cooling rate) critical; not sufficient alone.