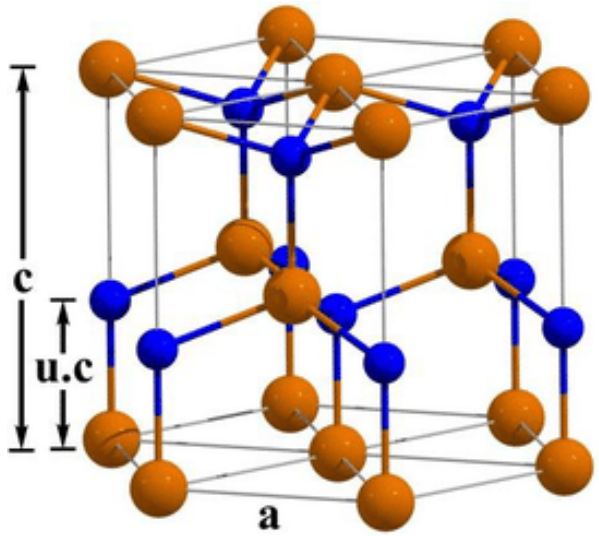
THE UNIVERSITY of TENNESSEE KNOXVILLE

# Lecture 10: ML for materials and ways to improve simple models

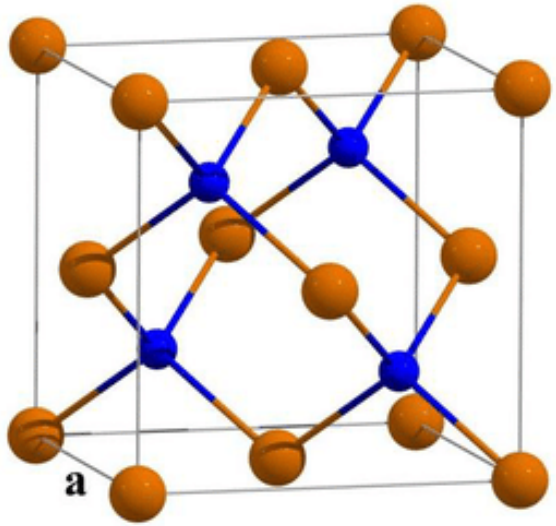
Instructor: Sergei V. Kalinin

# Binary Octet Compounds

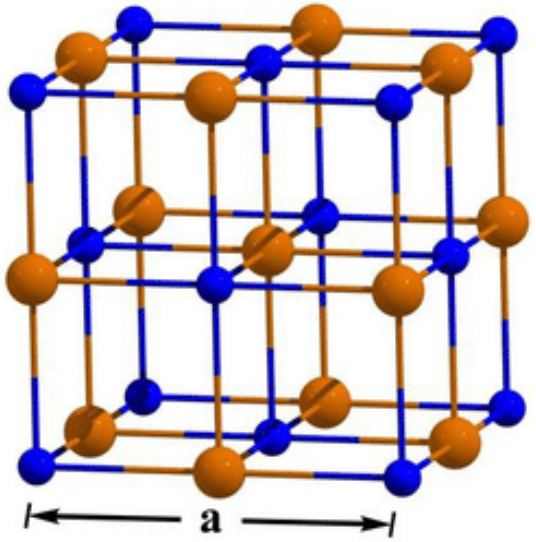
- NaCl, LiI, BeO, AlN, ….
- Can exist in  zincblende (ZB), wurtzite (WZ), rocksalt (RS), cesium chloride (CsCl), and diamond cubic (DC) crystal structures




(a) wurtzite



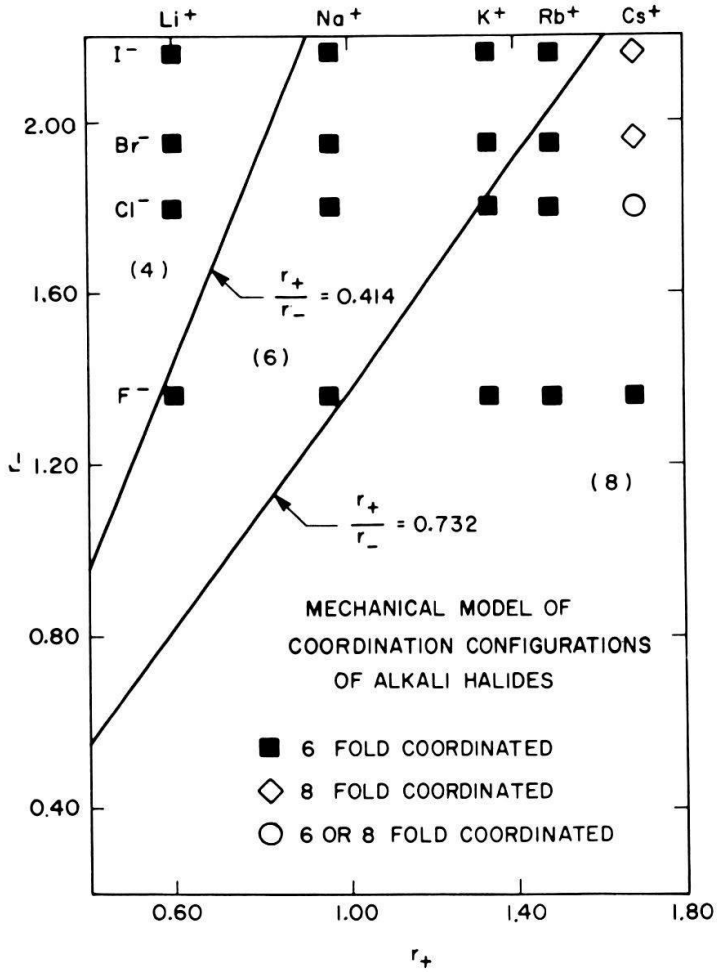

(b) zinc-blende



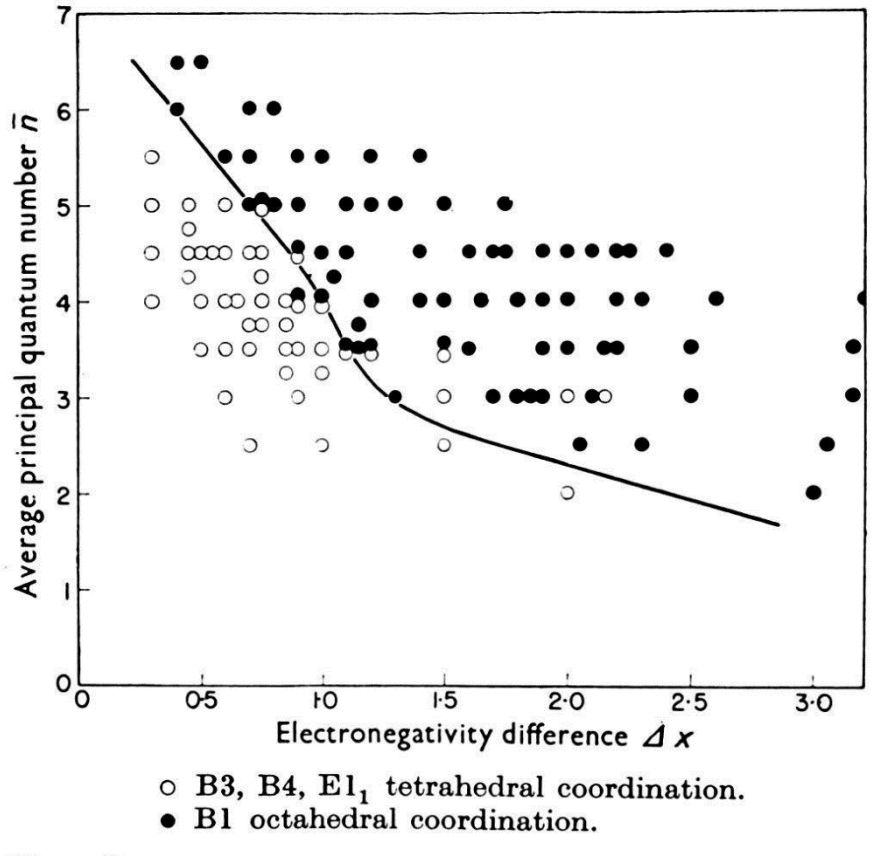

(c) rock-salt

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

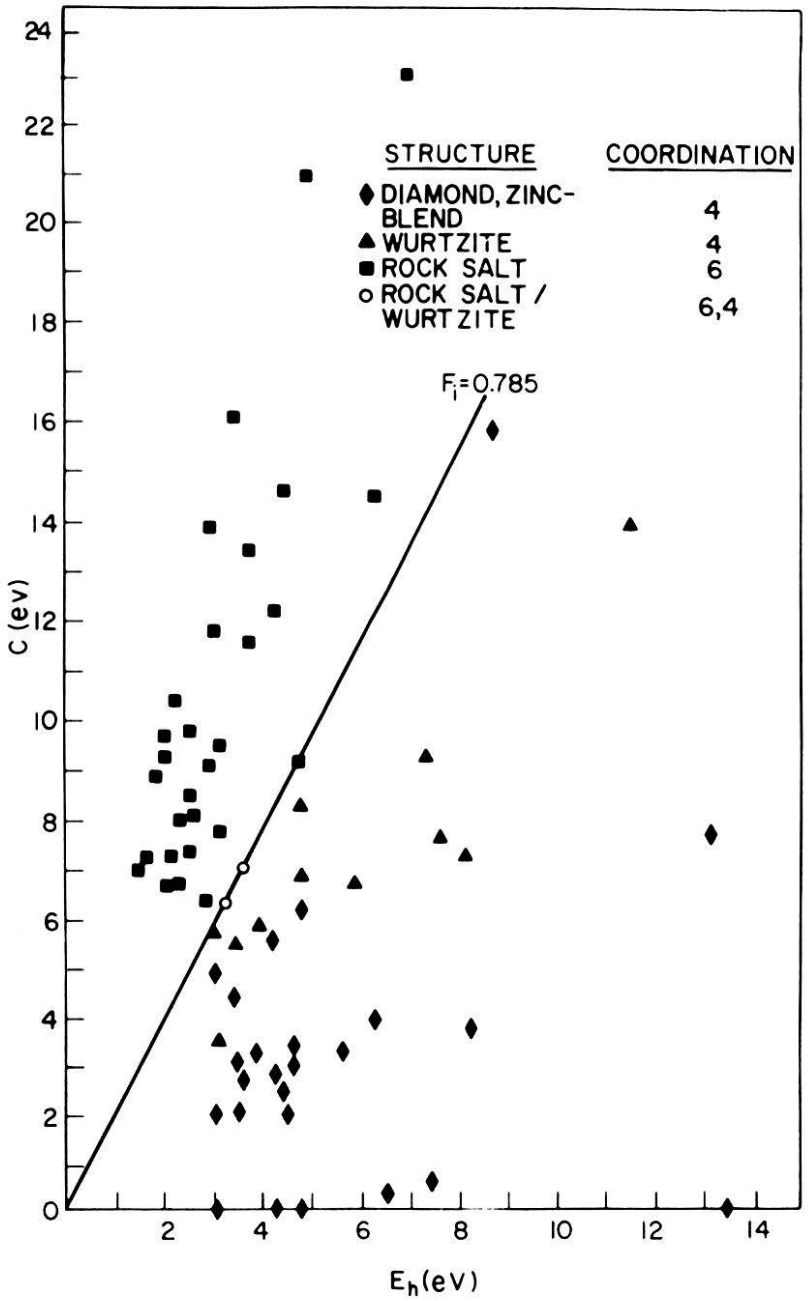
# Can we predict the structure from composition?




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 $\epsilon$. 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)

# Zunger diagrams

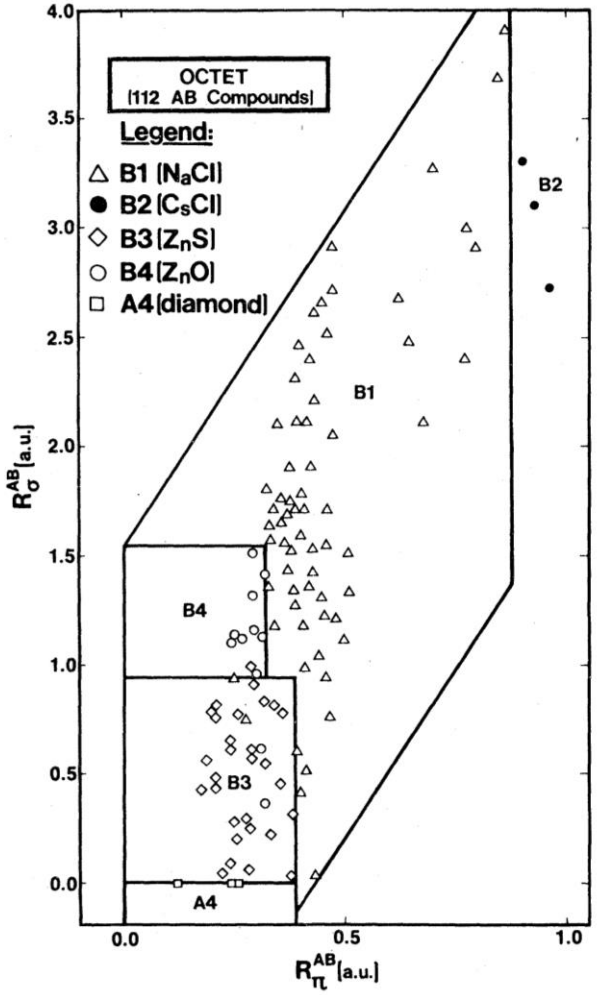



FIG. 19. Structural separation plot for the 112 binary octet compounds $A^N B^{(8-N)}$, obtained with the density-functional orbital radii, with

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

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

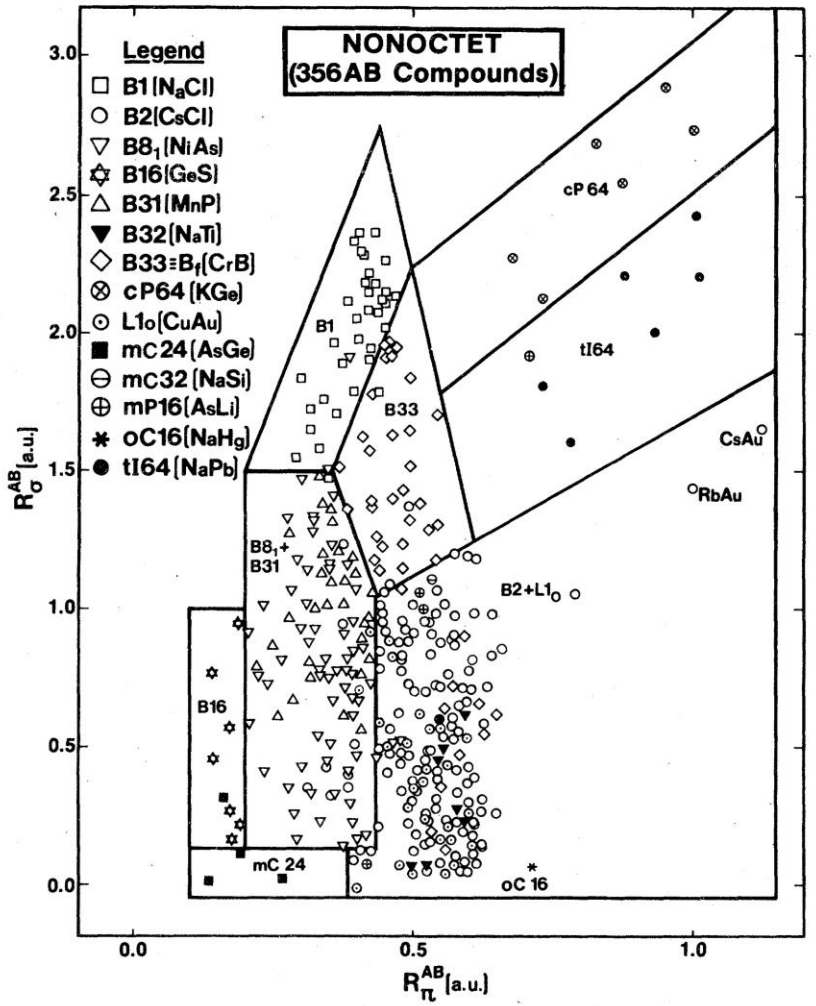



FIG. 20. Structural separation plot for the 356 binary nonoctet compounds, obtained with the density-functional orbital radii, with $R_\sigma^{AB} = |\,(r_p^A + r_s^A) - (r_p^B + r_s^B)\,|$, $R_\pi^{AB} = |\,r_p^A - r_s^A\,| + |\,r_p^B - r_s^B\,|$.

A. Zunger, Systematization of the stable crystal structure of all AB-type binary compounds: A pseudopotential orbital-radii approach, Phys. Rev. B 8, 15 (1980).

# Villars diagrams

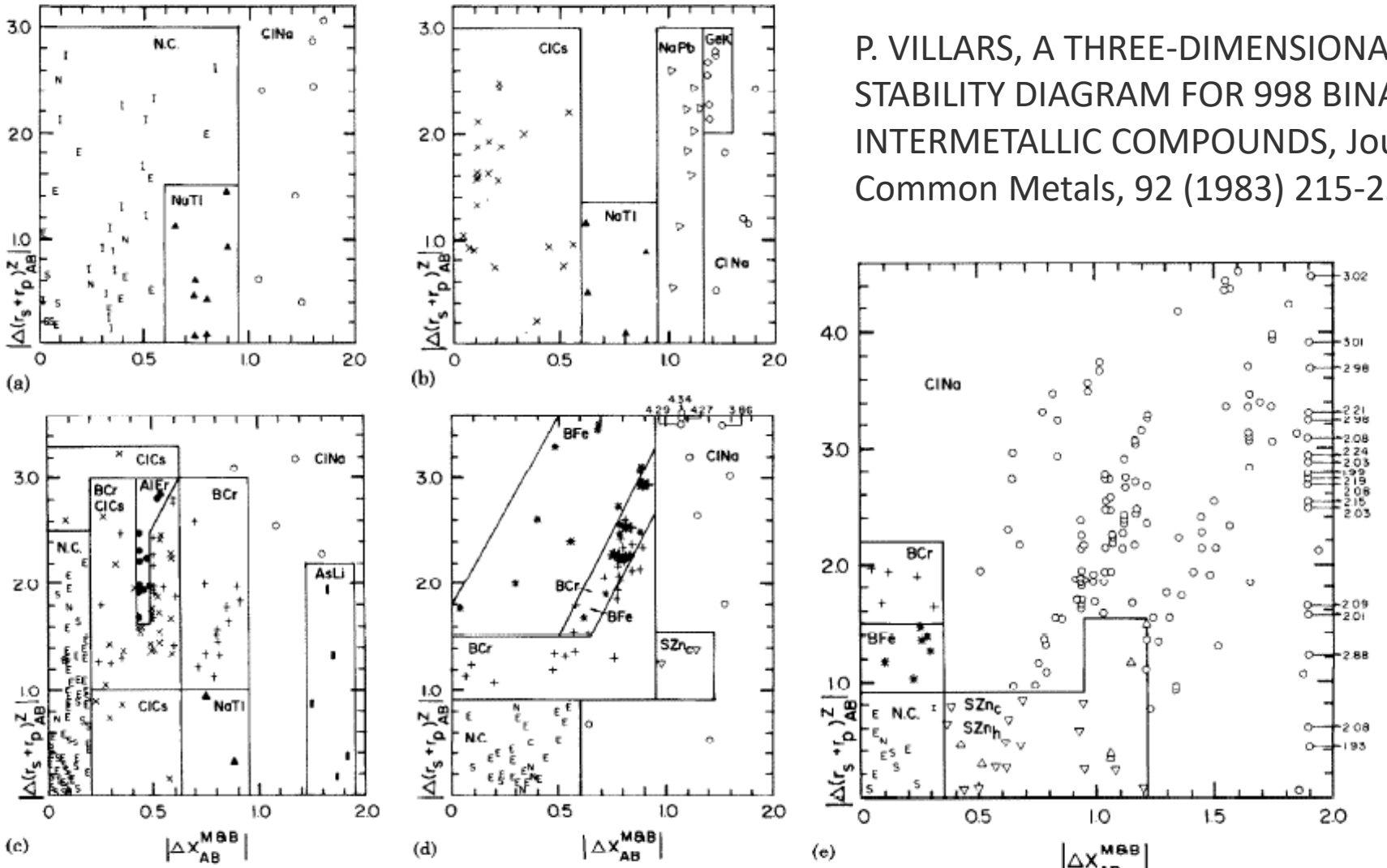
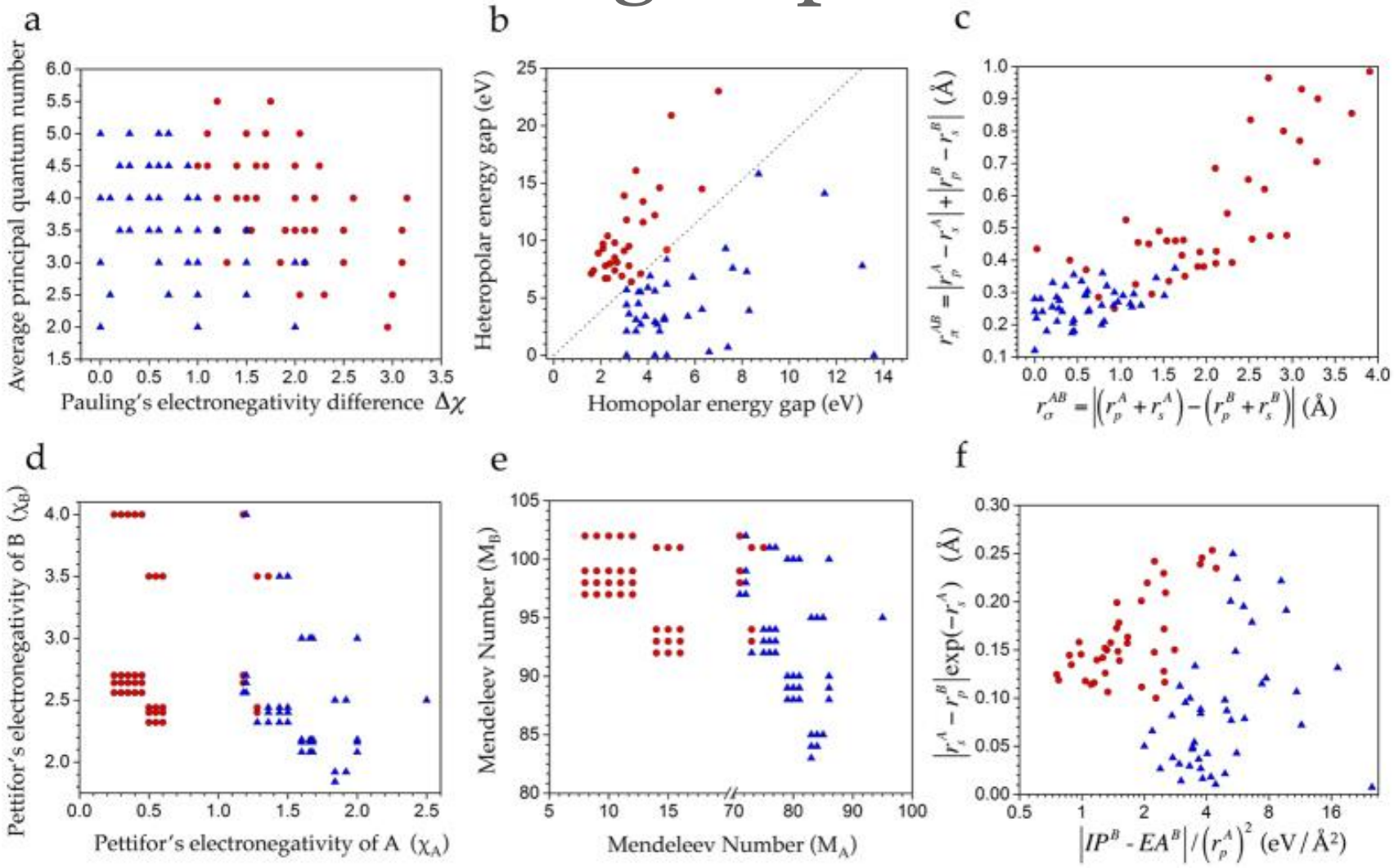


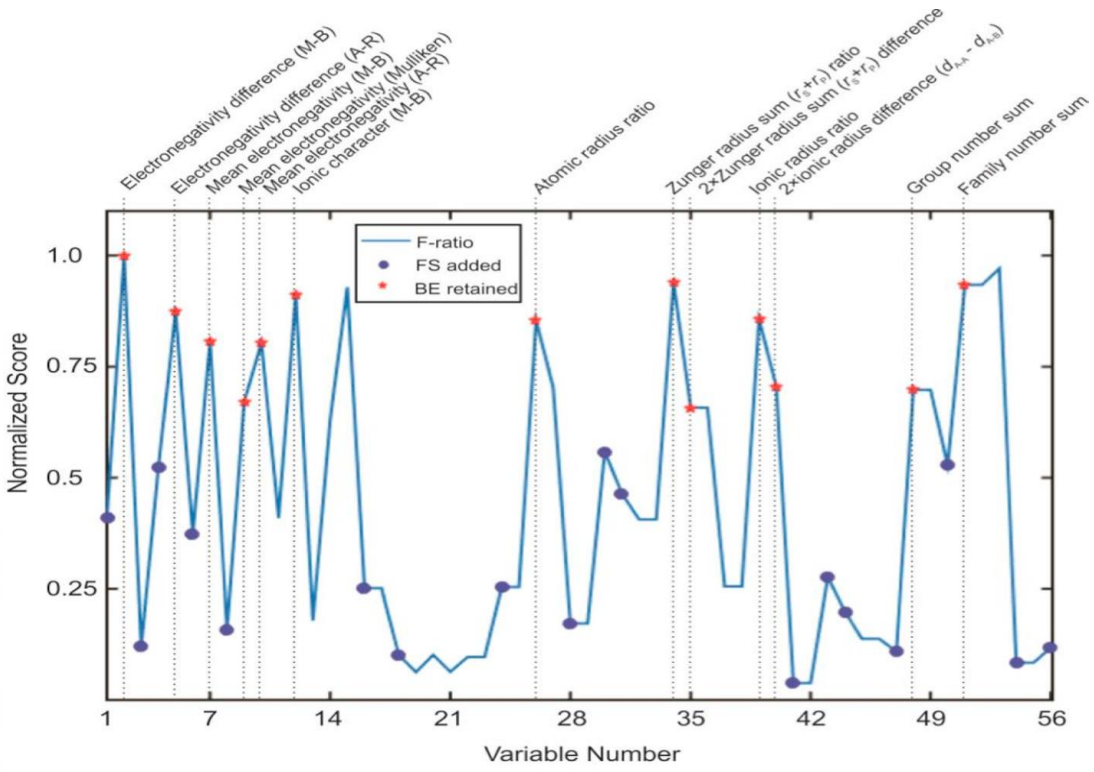

Fig. 3 *(continued).*

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

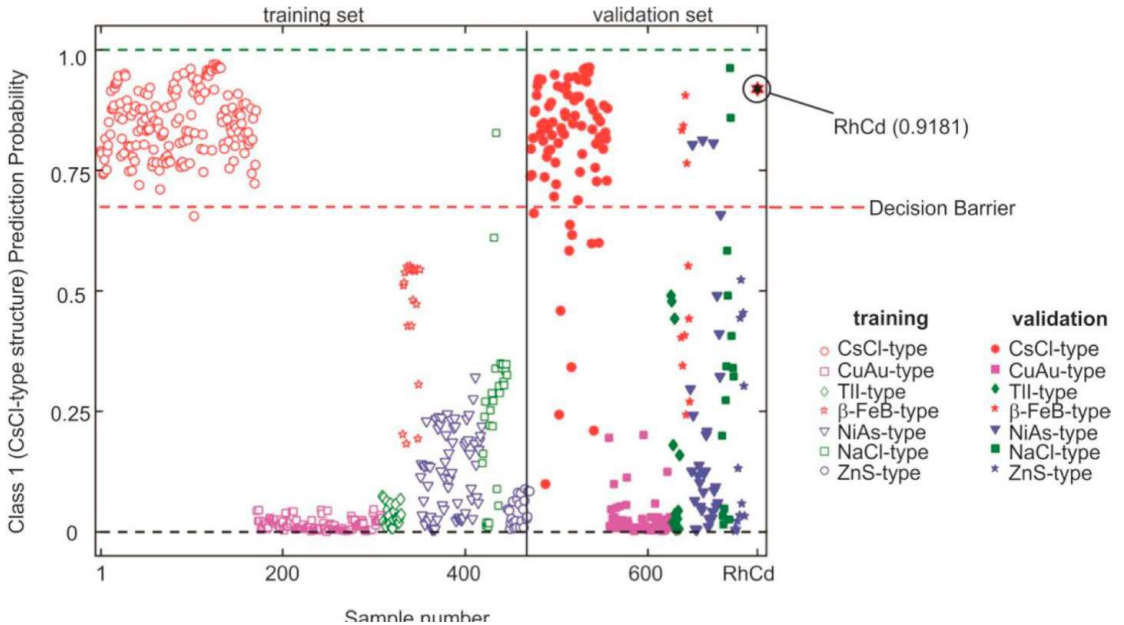
# Can machine learning help?




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

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

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

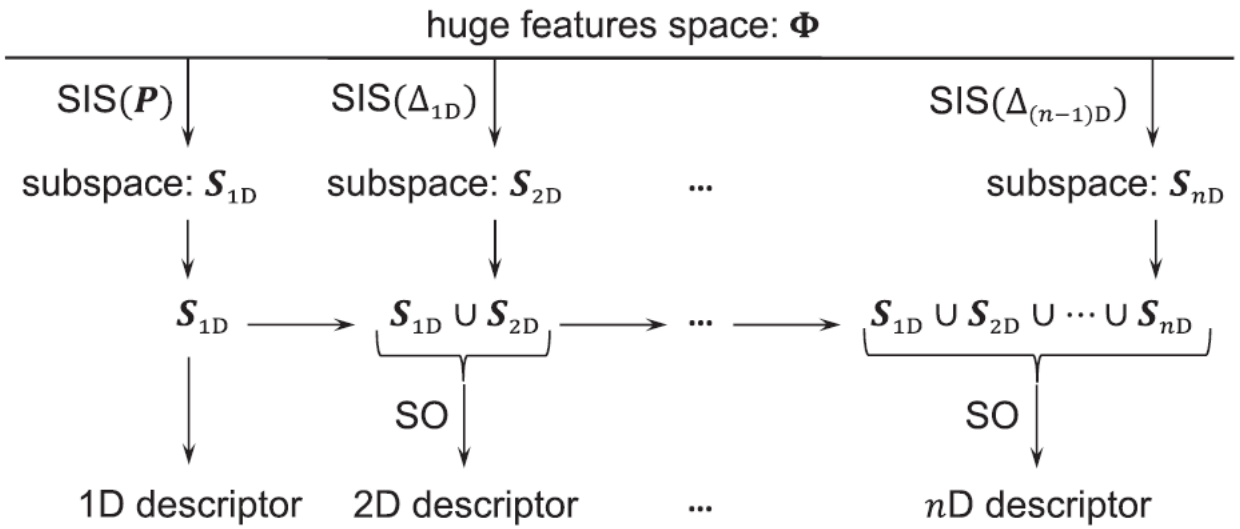



FIG. 1. The method SISSO combines unified subspaces having the largest correlation with residual errors $\boldsymbol{\Delta}$ (or $\boldsymbol{P}$) generated by sure independence screening (SIS) with sparsifying operator (SO) to further extract the best descriptor.

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 | $IE_A$, $IE_B$, $\chi_A$, $\chi_B$, $r_{\text{covA}}$, $r_{\text{covB}}$, $EA_A$, $EA_B$, $v_A$, $v_B$, $d_{AB}$ | $d_1 := \frac{IE_A IE_B (d_{AB} - r_{\text{covA}})}{\exp(\chi_A)\sqrt{r_{\text{covB}}}}$ | 100% |
| NaCl, CsCl, ZnS, $CaF_2$, $Cr_3Si$ | 217 | $IE_A$, $IE_B$, $\chi_A$, $\chi_B$, $r_{\text{covA}}$, $r_{\text{covB}}$, $d_{AB}$, $CN_A$, $CN_B$ | $d_1 := \frac{IE_B d_{AB}^2}{\chi_A r_{\text{covA}}^2 \sqrt{CN_B}}$, $d_2 := \frac{IE_A^2 r_{\text{covB}} \log(IE_A) \lvert r_{\text{covA}} - r_{\text{covB}} \rvert}{CN_B}$ | 100% |
| NaCl, CsCl, ZnS, $CaF_2$, $Cr_3Si$, SiC, $TiO_2$, ZnO, FeAs, NiAs | 260 | $IE_A$, $IE_B$, $\chi_A$, $\chi_B$, $r_{\text{covA}}$, $r_{\text{covB}}$, $d_{AB}$, $CN_A$, $CN_B$ | $d_1 := \frac{d_{AB}/r_{\text{covA}} - \chi_A/\chi_B}{\exp(CN_B/IE_B)}$, $d_2 := \frac{r_{\text{covA}}^3 d_{AB} IE_B}{\lvert \chi_B/\chi_A - \lvert CN_B - CN_A \rvert \rvert}$ | 99.6%[a] |
| NaCl, CsCl, ZnS, $CaF_2$, $Cr_3Si$, SiC, $TiO_2$, ZnO, FeAs, NiAs | 260 | $IE_A$, $IE_B$, $\chi_A$, $\chi_B$, $x_A$, $x_B$, $V_{\text{cell}}/\sum V_{\text{atom}}$ | $d_1 := \frac{V_{\text{cell}}}{\sum V_{\text{atom}}} \frac{\sqrt{\chi_B}}{\chi_A}$, $d_2 := \frac{IE_A IE_B}{\exp(V_{\text{cell}}/\sum V_{\text{atom}})}$ | 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 | $IE_A$, $IE_B$, $\chi_A$, $\chi_B$, $x_A$, $x_B$, $V_{\text{cell}}/\sum V_{\text{atom}}$ | $d_1 := \frac{x_B}{\sum V_{\text{atom}}/V_{\text{cell}}} \frac{IE_B \sqrt{\chi_B}}{\chi_A}$, $d_2 := \chi_A^2 \lvert \lvert 1 - 2x_A \rvert - x_A^2 \frac{\chi_B}{\chi_A} \rvert$ | 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

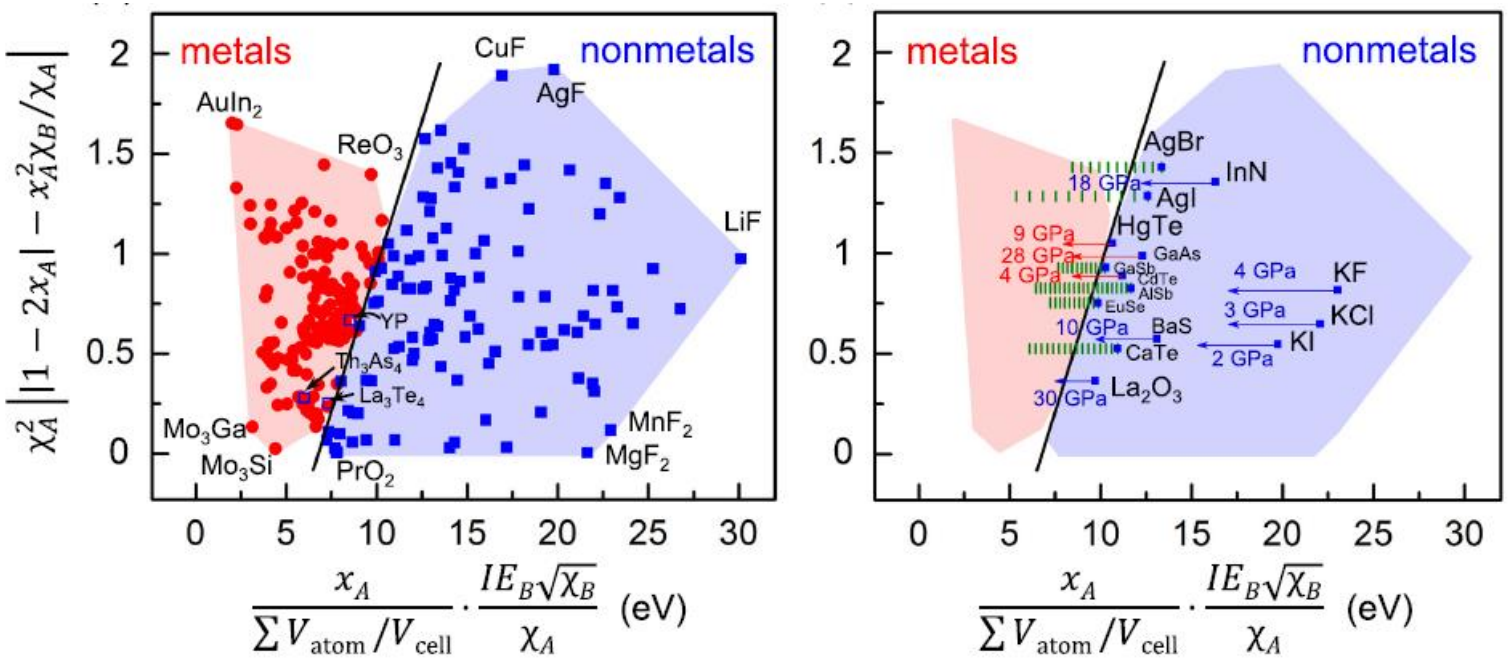
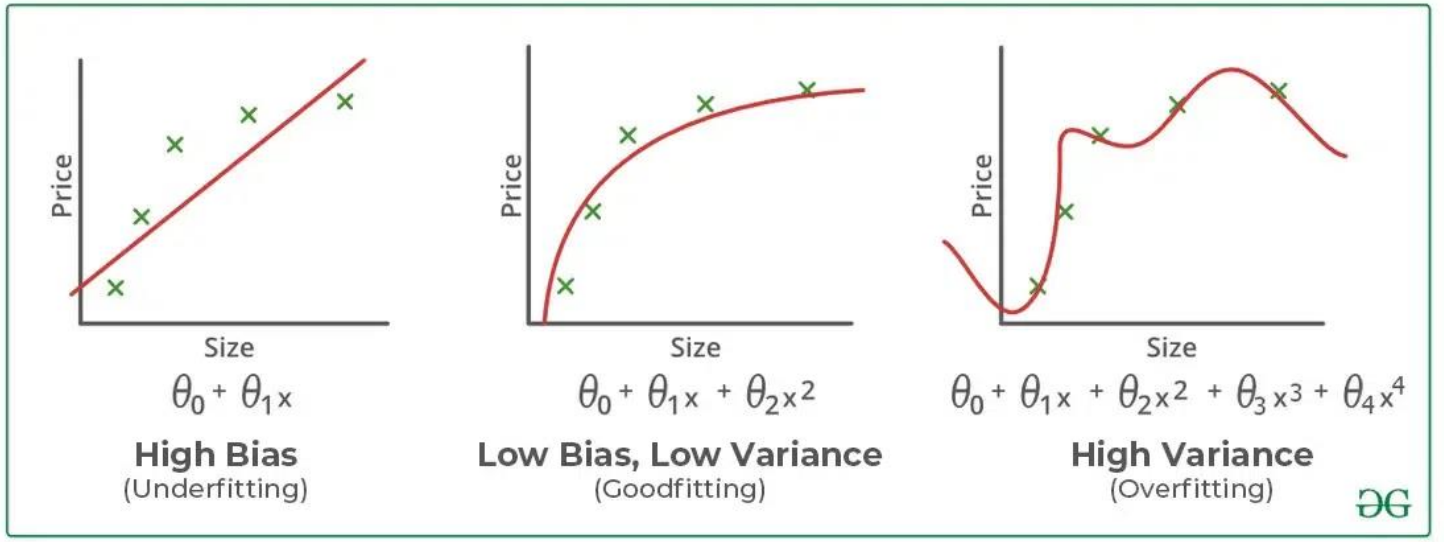
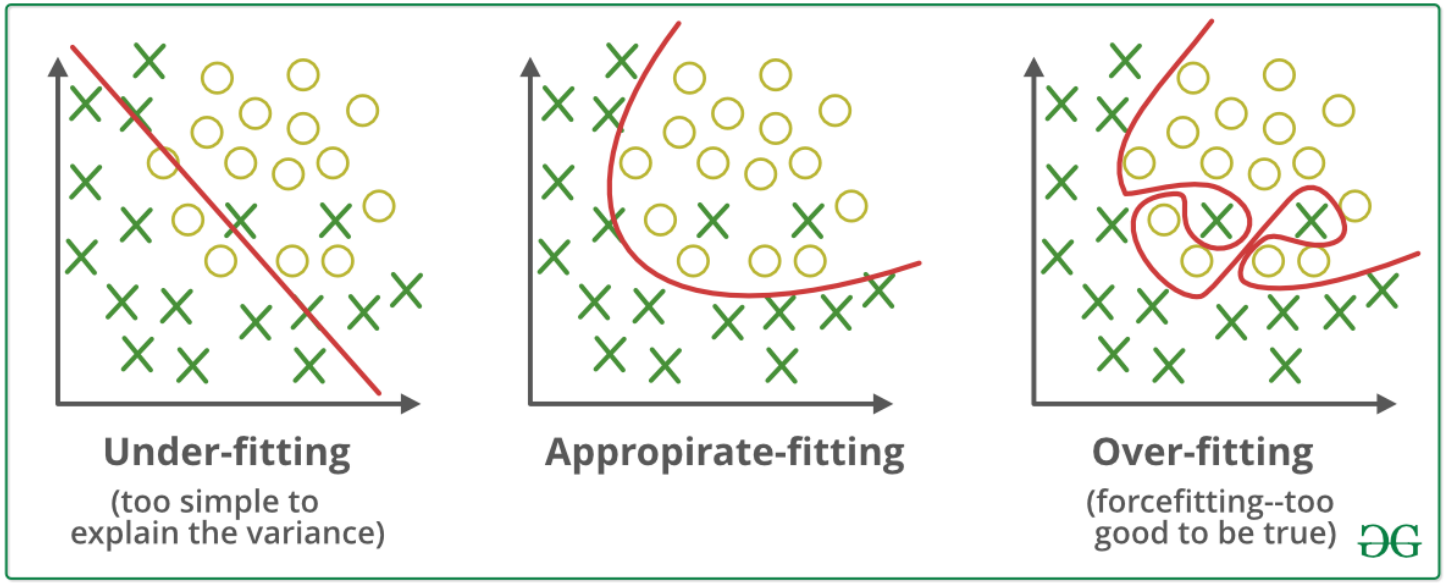



FIG. 4. SISSO for classification. (a) An almost perfect classification (99%) of metal/nonmetal for 299 materials. Symbols: $\chi$, Pauling electronegativity; $IE$, ionization energy; $x$, atomic composition; $\sum V_{\text{atom}}/V_{\text{cell}}$, packing fraction. Red circles, blue squares, and open 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)

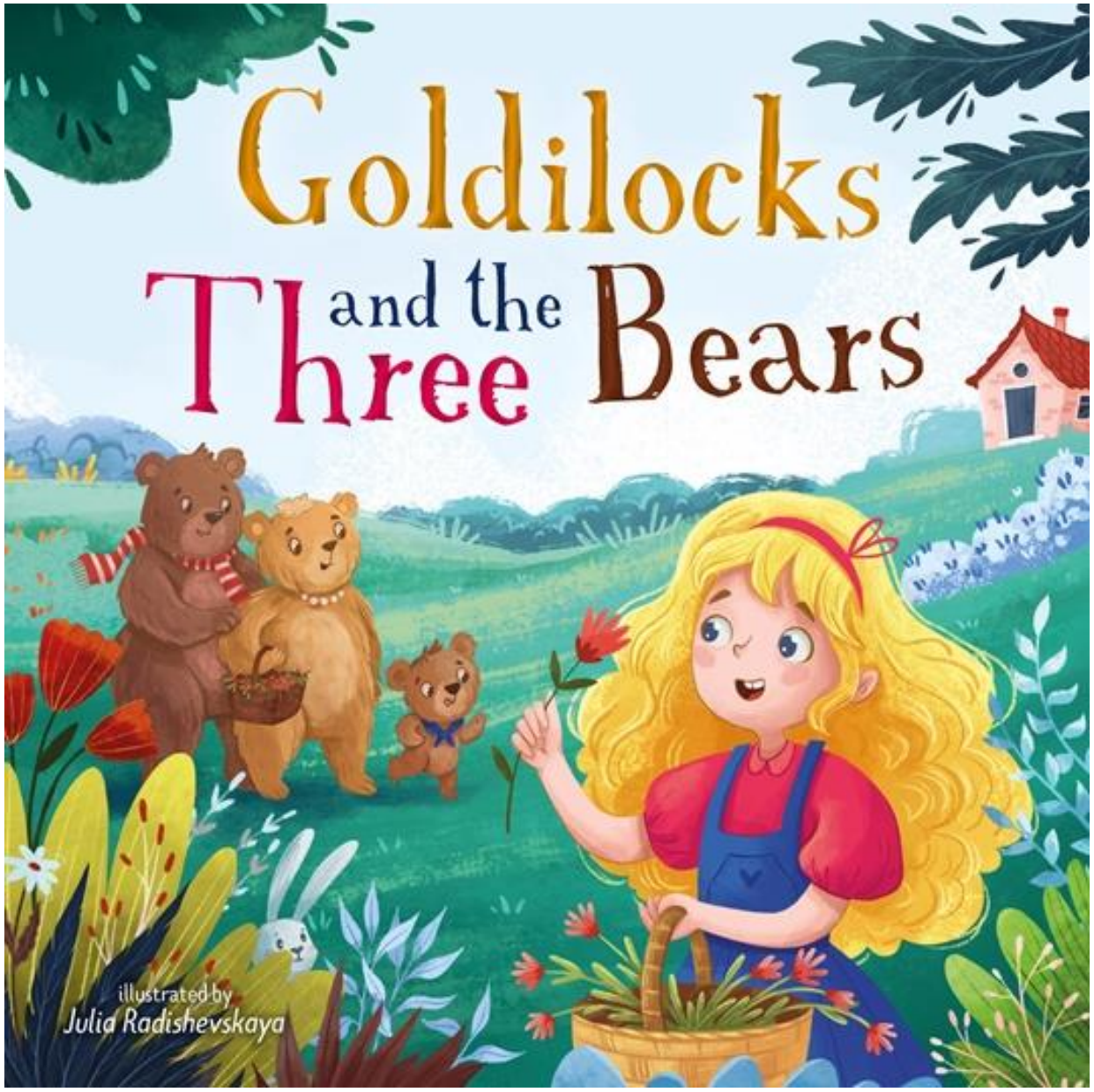
# Overfitting and Underfitting



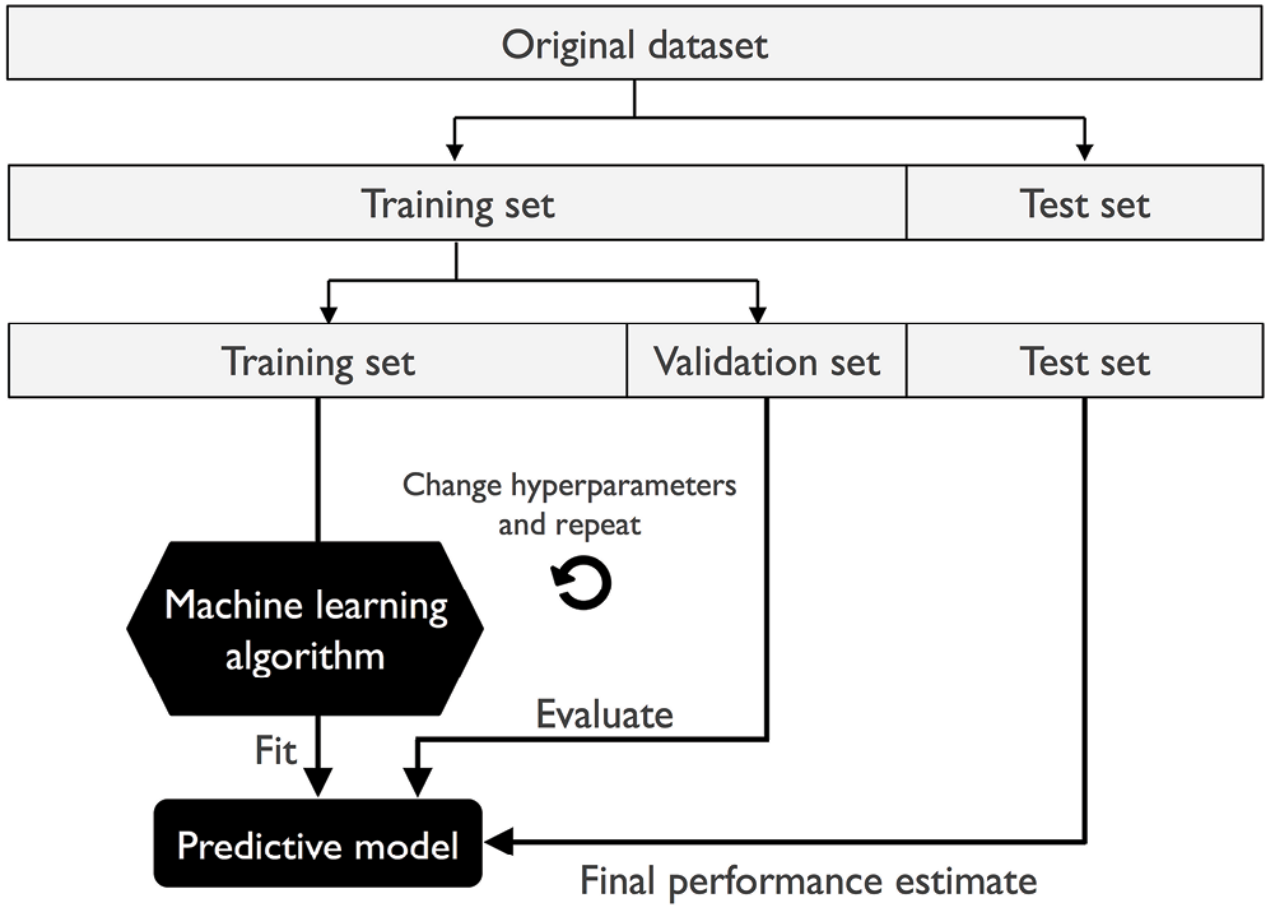

https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

# Overfitting and Underfitting



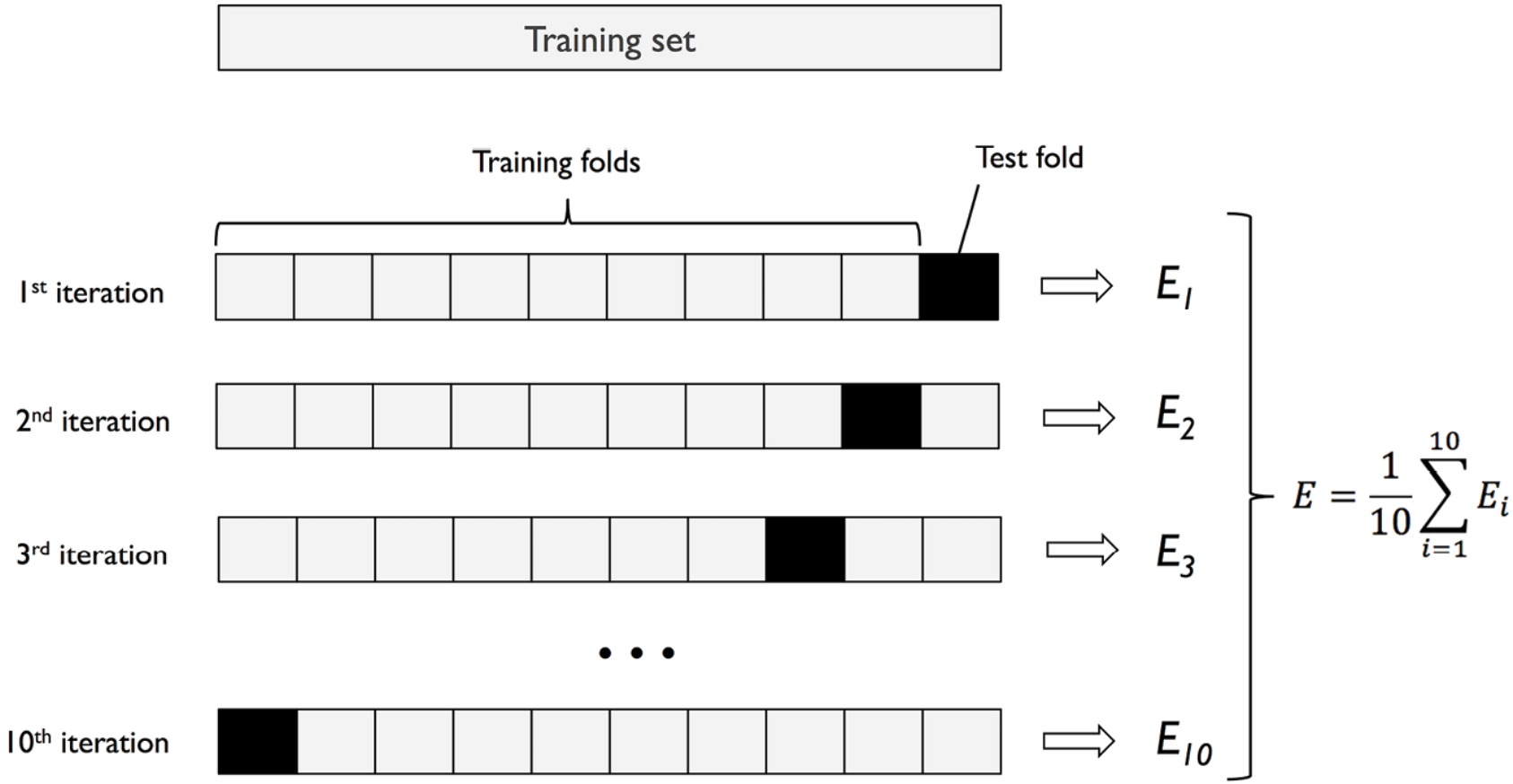
# Training, testing, and validating

How can we be certain that model that is trained on data we have will perform well in production?




From S. Raschka, Machine Learning with PyTorch and Scikit-Learn

# k-Fold cross-validation




From S. Raschka, Machine Learning with PyTorch and Scikit-Learn