

Advanced Materials Modeling:

Exploring Materials Space

*Center for Energy Science and Technology (CEST)
Skolkovo Institute of Science and Technology
Moscow, Russia*

High-throughput computational materials design

Top-down design:

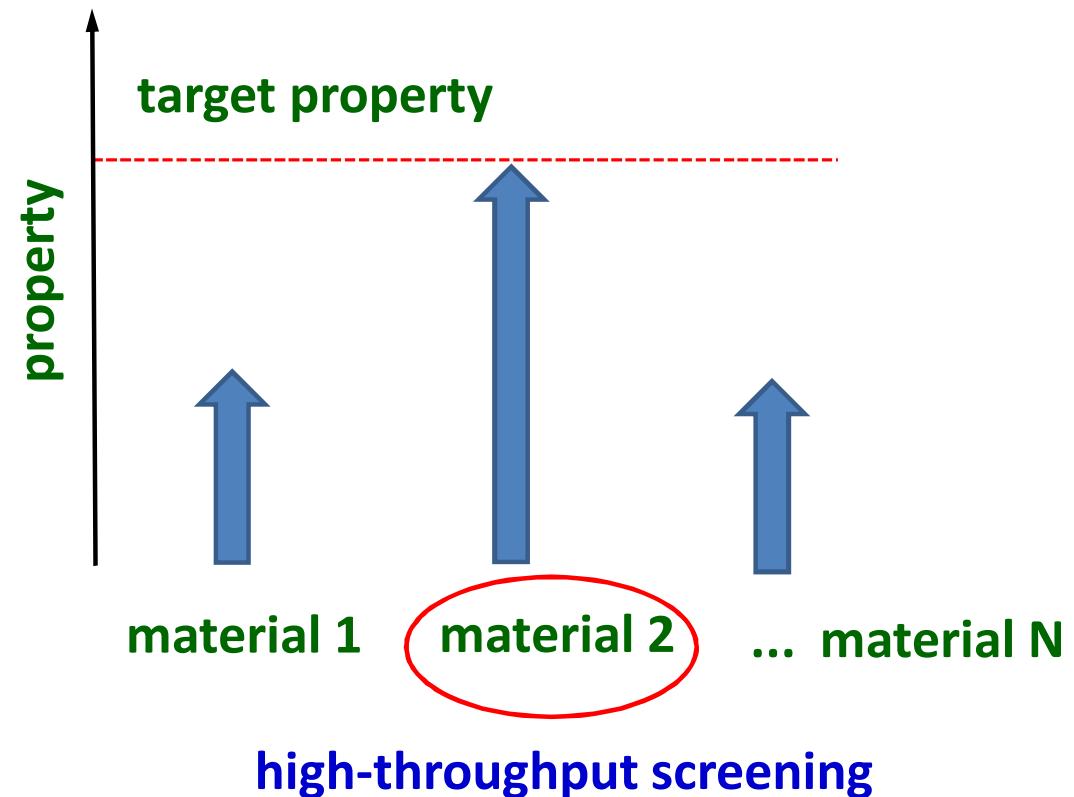
target property (high activity
and selectivity of a catalyst)

additional constraints
(high stability, low toxicity,...)

synthesis recipe

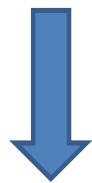
not clear how to achieve this!

Bottom-up design:



The key issue: Complexity

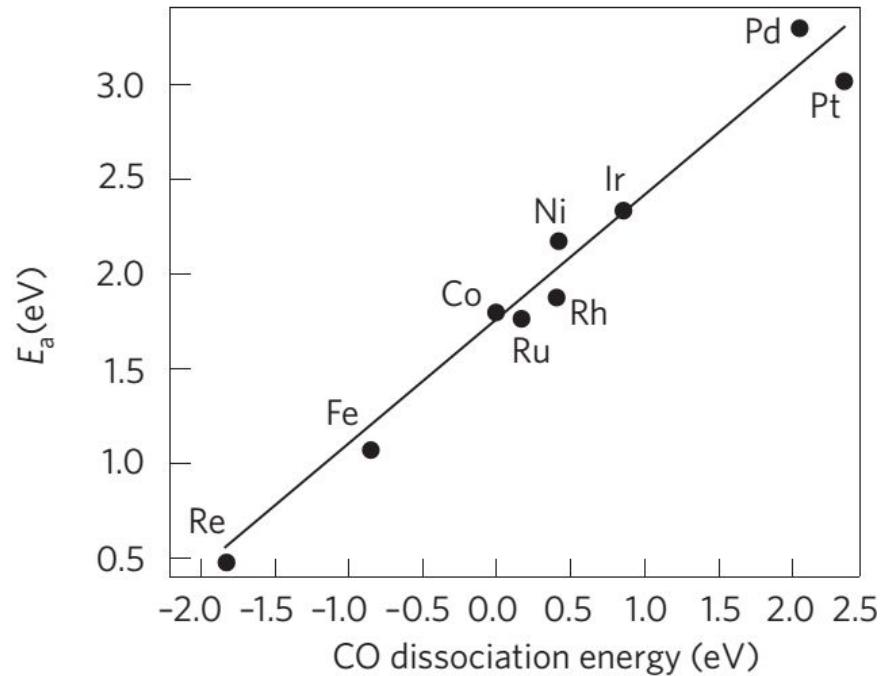
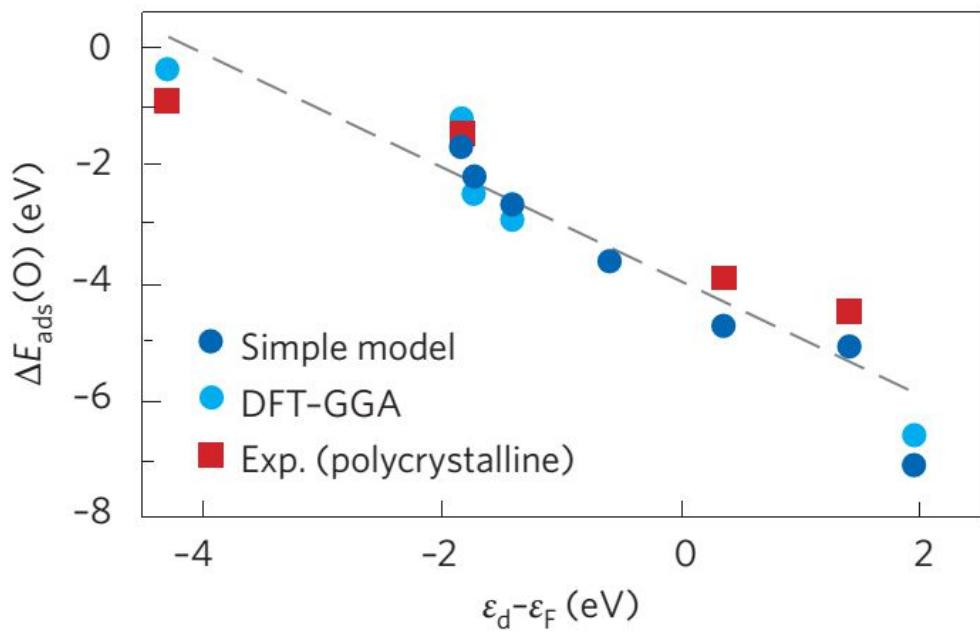
$$i \frac{\partial \Psi(x_1, x_2, \dots, x_n, R_1, R_2, \dots, R_N, t)}{\partial t} = \hat{H}(t) \Psi(x_1, x_2, \dots, x_n, R_1, R_2, \dots, R_N, t)$$



- 1) Many-body problem (3($n + N$)-dimensional)
- 2) Multiscale problem (tens orders of magnitude in time and space)

However, there is hope that the complexity can be treated *incrementally*

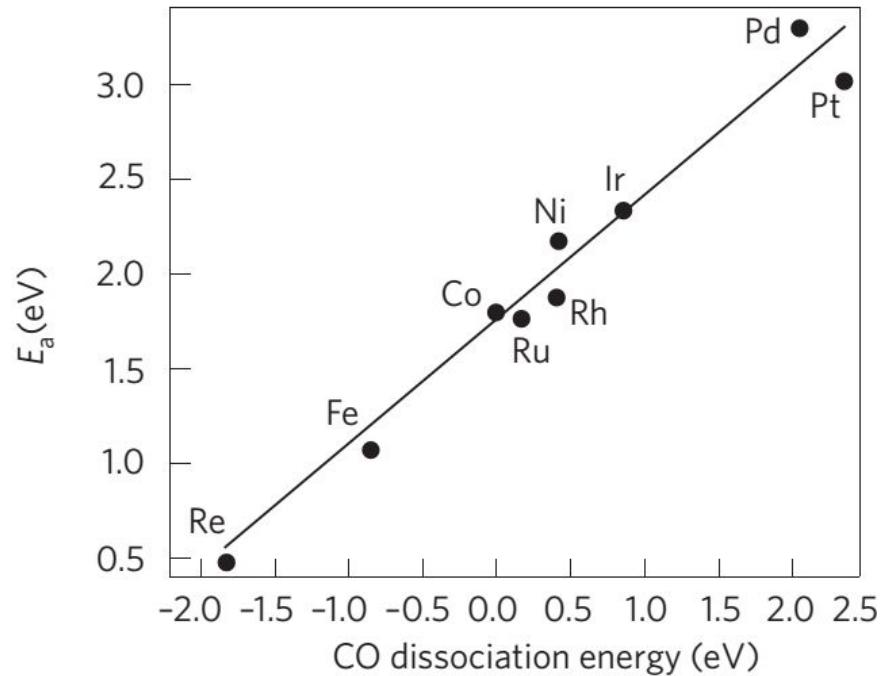
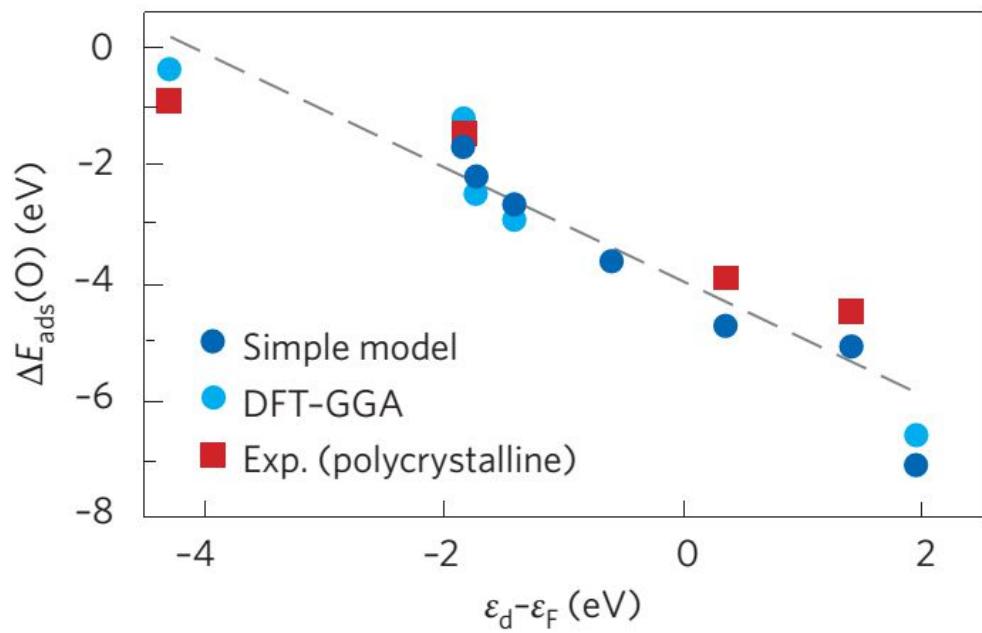
Descriptors



Simple(r) properties (bulk d-band center position and CO dissociation energy) are correlated to more complex properties (adsorption energy and reaction barrier)

The simpler quantities are called *descriptive parameters (a descriptor)*

Descriptors

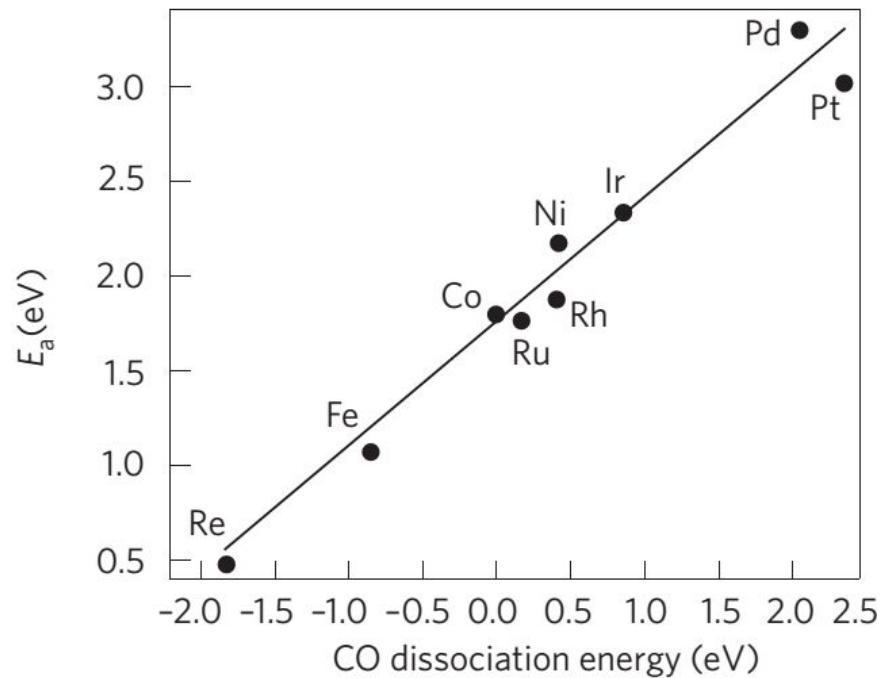
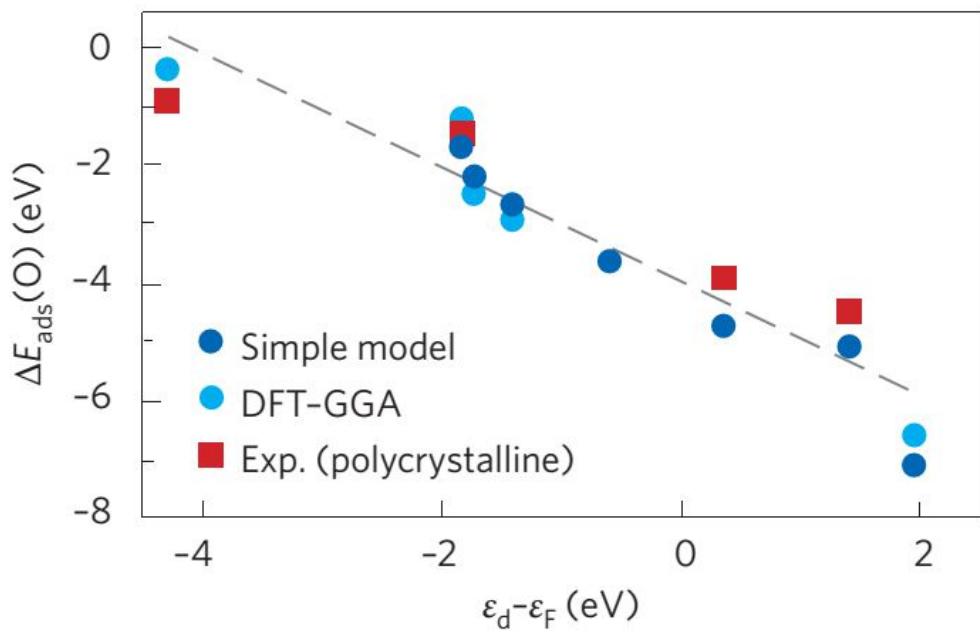


A simple physical model (Newns-Anderson) motivates the *d*-band center descriptor

What if we don't know such a model, or we need a more accurate and more widely applicable model?

J. K. Nørskov, T. Bligaard, J. Rossmeisl and C. H. Christensen, Nature Chemistry **1**, 37 (2009)

Descriptors

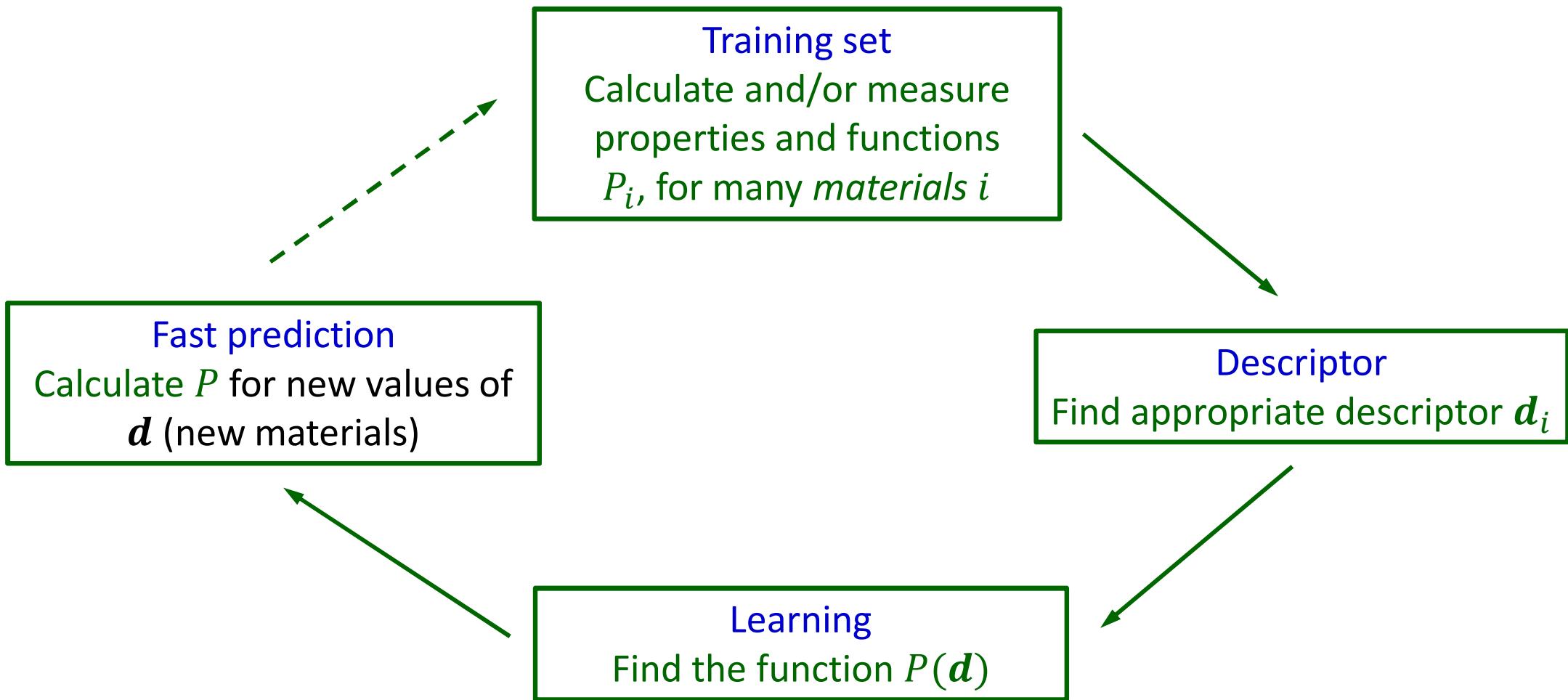


A simple physical model (Newns-Anderson) motivates the *d*-band center descriptor

Find descriptor from DATA!

J. K. Nørskov, T. Bligaard, J. Rossmeisl and C. H. Christensen, Nature Chemistry **1**, 37 (2009)

Supervised data analysis



Descriptors

- 1) A descriptor d_i uniquely characterizes the material i as well as property-relevant elementary processes**
- 2) The determination of the descriptor must not involve calculations as intensive as those needed for the evaluation of the property to be predicted**

Target property model: Kernel ridge regression versus feature selection

Regression models: Basis set expansion in materials space

kernel ridge regression

$$P(\mathbf{d}) = \sum_{i=1}^N c_i \exp(-\|\mathbf{d}_i - \mathbf{d}\|_2^2 / 2\sigma^2)$$

linear

$$P(\mathbf{d}) = \mathbf{d}\mathbf{c}$$

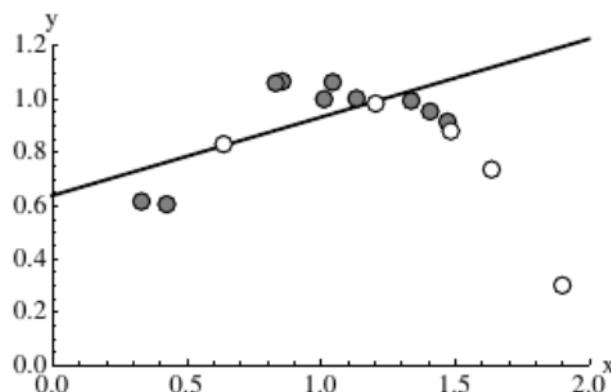
minimize

$$\|\mathbf{d}_i - \mathbf{d}_j\|_2^2 = \sum_{\alpha=1}^{\Omega} (d_{i,\alpha} - d_{j,\alpha})^2$$

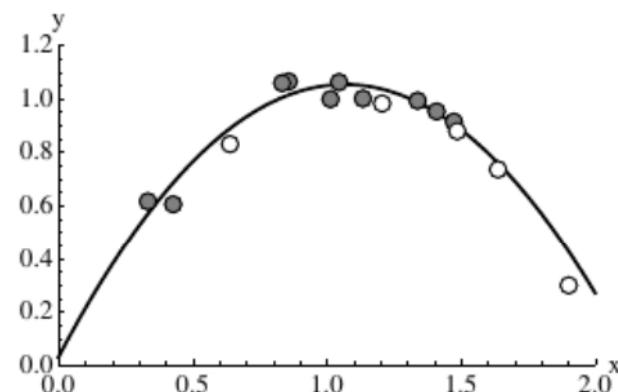
Regression: Importance of regularization

● training ○ validation

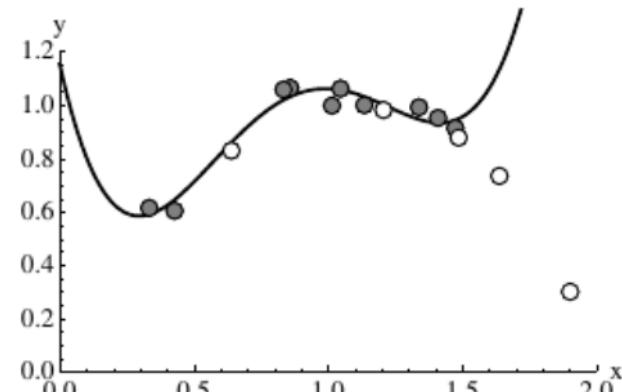
Underfitting



Fitting



Overfitting



Training/
validation
error

0.044 / 0.068

0.036 / 0.939

$$\min_c \sum_i (P(d_i, c) - P_i)^2 + \lambda f(c), \min_{\lambda} (\text{validation error}) \rightarrow \lambda$$

Target property model: Kernel ridge regression versus feature selection

Regression models: Basis set expansion in materials space

kernel ridge regression

$$P(\mathbf{d}) = \sum_{i=1}^N c_i \exp(-\|\mathbf{d}_i - \mathbf{d}\|_2^2 / 2\sigma^2)$$

$$\sum_{i=1}^N (P(\mathbf{d}_i) - P_{\textcolor{blue}{i}})^2 +$$

$$\lambda \sum_{i,j=1}^{N,N} c_i c_j \exp(-\|\mathbf{d}_i - \mathbf{d}_j\|_2^2 / 2\sigma^2)$$

$$\|\mathbf{d}_i - \mathbf{d}_j\|_2^2 = \sum_{\alpha=1}^{\Omega} (d_{i,\alpha} - d_{j,\alpha})^2$$

linear

$$P(\mathbf{d}) = \mathbf{d}\mathbf{c}$$

minimize

$$\sum_{i=1}^N (P(\mathbf{d}_i) - P_{\textcolor{blue}{i}})^2 + \lambda \|\mathbf{c}\|_0$$

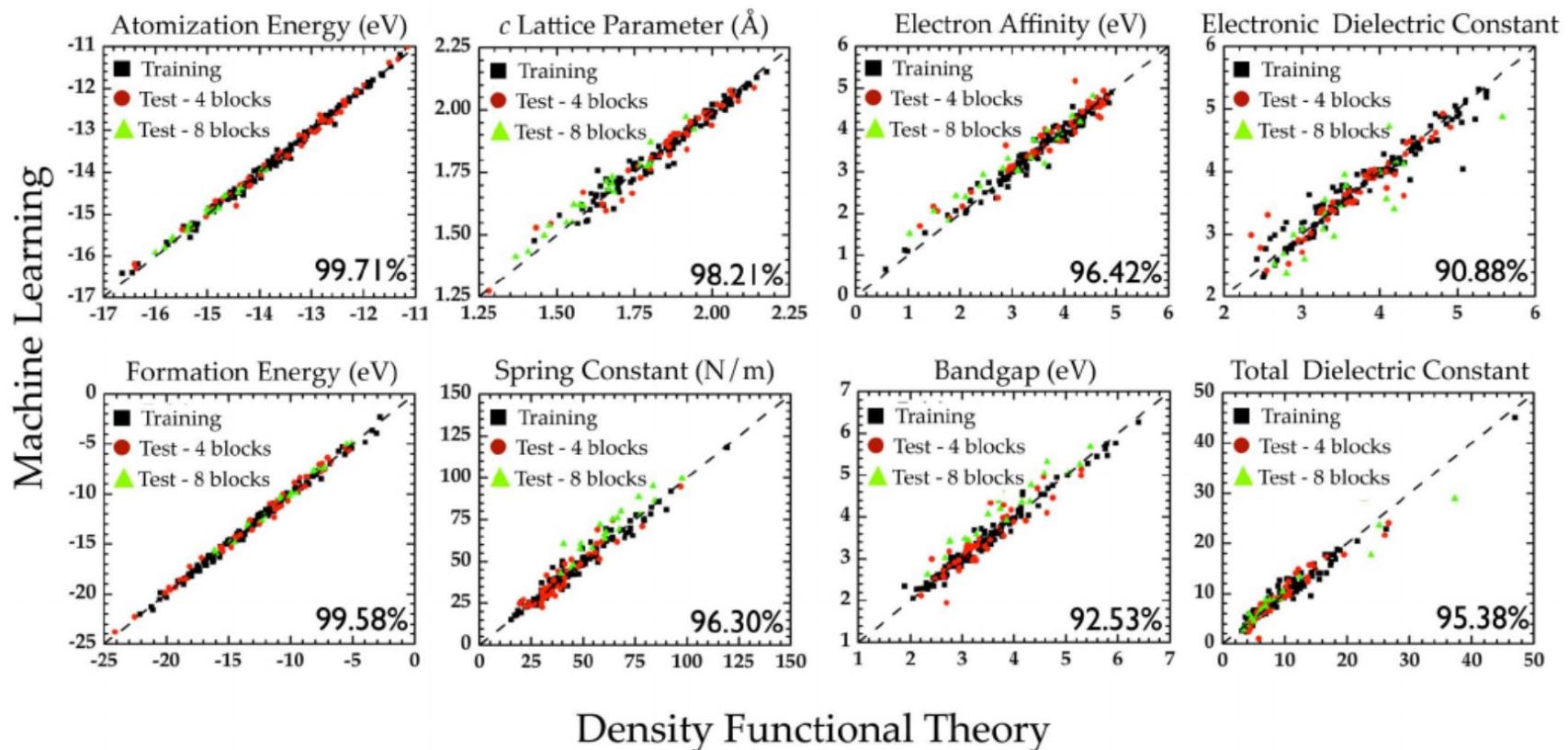
Target property model: Kernel ridge regression versus feature selection

| kernel (Gaussian, Laplacian, linear ($d_i \cdot d_j$)) | |
|--|---|
| kernel ridge regression | linear |
| $P(\mathbf{d}) = \sum_{i=1}^N c_i \exp(-\ \mathbf{d}_i - \mathbf{d}\ _2^2 / 2\sigma^2)$ | $P(\mathbf{d}) = \mathbf{d}\mathbf{c}$ |
| $\sum_{i=1}^N (P(\mathbf{d}_i) - P_{\mathbf{i}})^2 +$ $\lambda \sum_{i,j=1}^{N,N} c_i c_j \exp(-\ \mathbf{d}_i - \mathbf{d}_j\ _2^2 / 2\sigma^2)$ | $\sum_{i=1}^N (P(\mathbf{d}_i) - P_{\mathbf{i}})^2 +$ $\lambda \ \mathbf{c}\ _0$ |
| minimize | penalty on the number of non-zero coefficients $\ \mathbf{c}\ _0$ |
| penalty on similar data points | |

(Gaussian) kernel ridge regression example

Data: 175 linear 4-blocks periodic polymers. 7 blocks: CH_2 , SiF_2 , SiCl_2 , GeF_2 ,
 GeCl_2 , SnF_2 , SnCl_2 ,

Descriptor: 20 dimensions [# building blocks of type i , of ii pairs, of iii triplets]



Descriptors

- 1) A descriptor d_i uniquely characterizes the material i as well as property-relevant elementary processes**
- 2) The determination of the descriptor must not involve calculations as intensive as those needed for the evaluation of the property to be predicted**
- 3) The dimension Ω of the descriptor should be as low as possible (for a certain accuracy request)**

Choose a physically motivated basis set!

Descriptors

- 1) A descriptor d_i uniquely characterizes the material i as well as property-relevant elementary processes
- 2) The determination of the descriptor must not involve calculations as intensive as those needed for the evaluation of the property to be predicted
- 3) The dimension Ω of the descriptor should be as low as possible (for a certain accuracy request)

Idea: calculate many *physically motivated* quantities (features), and use these features as a basis for the physical model under compactness constraints

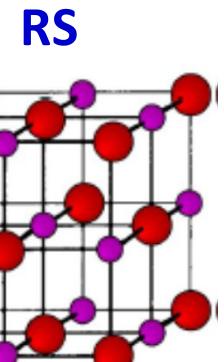
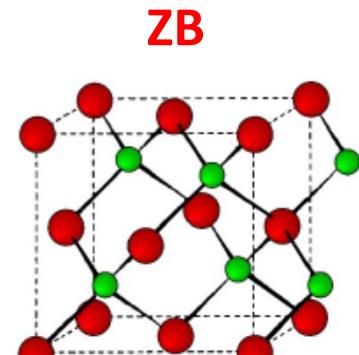
Proof of Concept: Descriptor for the Classification “Zincblende/Wurtzite (ZB/W) or Rocksalt (RS)?”

82 binary octet compounds

| | | |
|--------------------|---------|-----------|
| hydrogen | 1 | |
| H | 1.0070 | |
| lithium | 3 | beryllium |
| Li | 6.941 | Be |
| sodium | 11 | magnesium |
| Na | 22.990 | Mg |
| potassium | 19 | calcium |
| K | 39.098 | Ca |
| rubidium | 37 | strontium |
| Rb | 85.468 | Sr |
| caesium | 55 | barium |
| Cs | 132.91 | Ba |
| francium | * | Lu |
| radium | -89-102 | Hf |
| copper | 29 | Ta |
| Zn | 63.546 | W |
| silver | 47 | Re |
| Ag | 65.39 | Os |
| cadmium | 48 | Ir |
| Cd | 138.91 | Pt |
| Lanthanide series | | Au |
| ** Actinide series | | Hg |

| | | | | | | | | | | | | | | | |
|--------------|--------|-----------|--------|----------|--------|------------|--------|------------|--------|-----------|--------|---------|--------|-----------|--------|
| scandium | 21 | titanium | 22 | vanadium | 23 | chromium | 24 | manganese | 25 | iron | 26 | cobalt | 27 | nickel | 28 |
| Sc | 44.956 | Ti | 47.867 | V | 50.942 | Cr | 51.996 | Mn | 54.938 | Fe | 55.845 | Co | 58.933 | Ni | 58.693 |
| yttrium | 39 | zirconium | 40 | niobium | 41 | molybdenum | 42 | technetium | 43 | ruthenium | 44 | rhodium | 45 | palladium | 46 |
| Y | 88.906 | Zr | 91.224 | Nb | 92.906 | 95.94 | [98] | Tc | 101.07 | Ru | 102.91 | Rh | 106.42 | | |
| lanthanum | 57 | lutetium | 71 | hafnium | 72 | tantalum | 73 | tungsten | 74 | rhenium | 75 | osmium | 76 | iridium | 77 |
| La | 138.91 | | | | | | | | | | | | | | |
| cerium | 58 | | | | | | | | | | | | | | |
| Ce | 140.12 | | | | | | | | | | | | | | |
| praseodymium | 59 | | | | | | | | | | | | | | |
| Pr | 140.91 | | | | | | | | | | | | | | |
| neodymium | 60 | | | | | | | | | | | | | | |
| Nd | 144.24 | | | | | | | | | | | | | | |
| promethium | 61 | | | | | | | | | | | | | | |
| Pm | [145] | | | | | | | | | | | | | | |
| samarium | 62 | | | | | | | | | | | | | | |
| Sm | 150.36 | | | | | | | | | | | | | | |
| europlum | 63 | | | | | | | | | | | | | | |
| Eu | 151.96 | | | | | | | | | | | | | | |
| gadolinium | 64 | | | | | | | | | | | | | | |
| Gd | 157.25 | | | | | | | | | | | | | | |
| terbium | 65 | | | | | | | | | | | | | | |
| Tb | 158.93 | | | | | | | | | | | | | | |
| dysprosium | 66 | | | | | | | | | | | | | | |
| Dy | 162.50 | | | | | | | | | | | | | | |
| holmium | 67 | | | | | | | | | | | | | | |
| Ho | 164.93 | | | | | | | | | | | | | | |
| erbium | 68 | | | | | | | | | | | | | | |
| Er | 167.26 | | | | | | | | | | | | | | |
| thulium | 69 | | | | | | | | | | | | | | |
| Tm | 168.93 | | | | | | | | | | | | | | |
| yterbium | 70 | | | | | | | | | | | | | | |
| Yb | 173.04 | | | | | | | | | | | | | | |

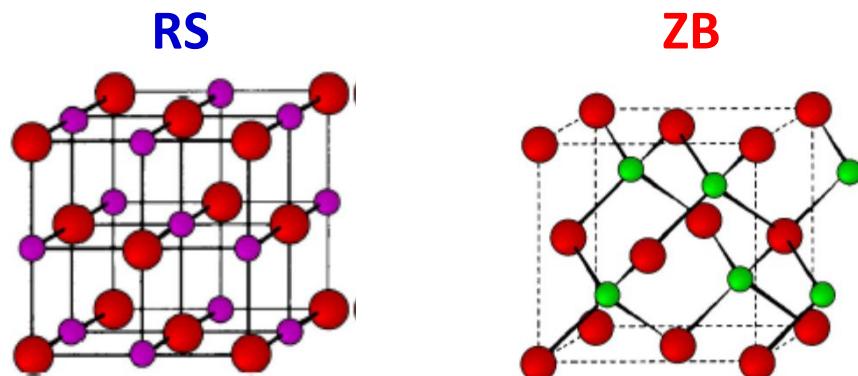
| | | | | | | | | | | | |
|-------------|--------|-----------|--------|------------|--------|----------|--------|----------|--------|---------|-------|
| boron | 5 | carbon | 6 | nitrogen | 7 | oxygen | 8 | fluorine | 9 | helium | 2 |
| B | 10.811 | C | 12.011 | N | 14.007 | O | 15.999 | F | 18.998 | neon | 10 |
| aluminum | 13 | silicon | 14 | phosphorus | 15 | sulfur | 16 | chlorine | 17 | argon | 18 |
| Al | 26.982 | Si | 28.086 | P | 30.974 | S | 32.065 | Br | 35.453 | krypton | 36 |
| gallium | 31 | germanium | 32 | arsenic | 33 | selenium | 34 | Iodine | 53 | xenon | 54 |
| Ga | 69.723 | Ge | 72.61 | As | 74.922 | Te | 78.96 | Po | 126.90 | radon | 86 |
| In | 114.82 | Sn | 118.71 | Sb | 121.76 | Bi | 127.60 | At | [210] | Rn | [222] |
| tin | 50 | antimony | 51 | bismuth | 83 | polonium | 84 | astatine | 85 | | |
| Tl | 204.38 | Pb | 207.2 | Bi | 208.98 | Po | [209] | At | [210] | | |
| ununquadium | 114 | Uuo | [289] | | | | | | | | |



Proof of Concept: Descriptor for the Classification “Zincblende/Wurtzite (ZB/W) or Rocksalt (RS)?”

Crystal-structure prediction was and is one of the most important, basic challenges of materials science and engineering.

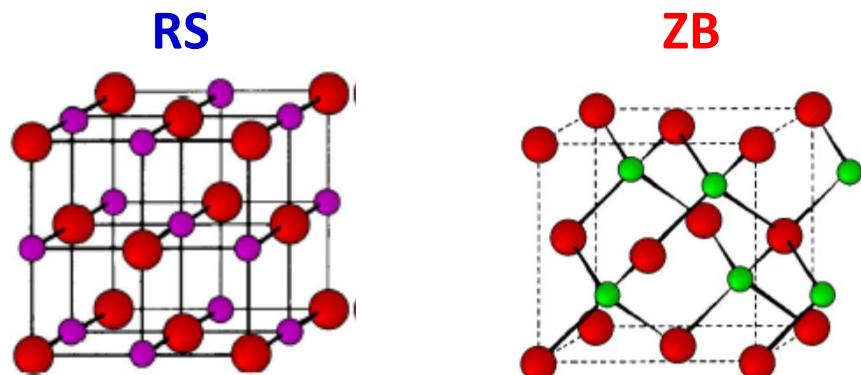
Energy differences between different structures are very small.



For Si: 0.01% of the energy of a Si atom, or 0.1% of the 4 valence electrons.

Proof of Concept: Descriptor for the Classification “Zincblende/Wurtzite (ZB/W) or Rocksalt (RS)?”

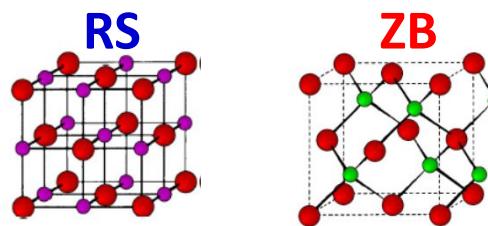
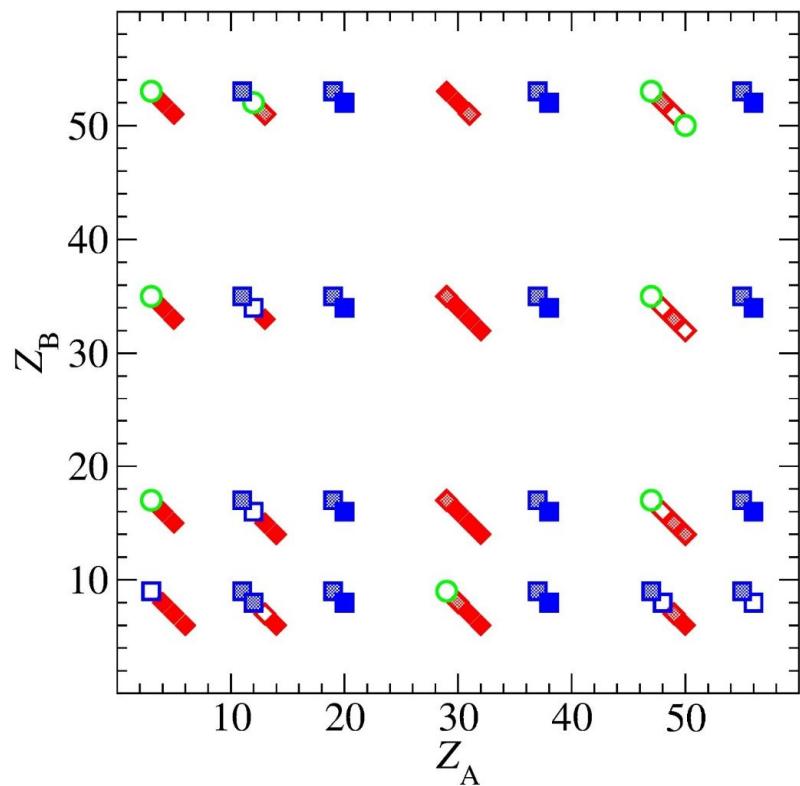
Crystal-structure prediction was and is one of the most important, basic challenges of materials science and engineering.



J. A. van Vechten, Phys. Rev. 182, 891 (1969). J. C. Phillips, Rev. Mod. Phys. 42, 317 (1970).
J. John and A.N. Bloch, Phys. Rev. Lett. 33, 1095 (1974) J. R. Chelikowsky and J. C. Phillips, Phys. Rev. B 33, 2453 (1978)
A. Zunger, Phys. Rev. B 22, 5839 (1980).
D. G. Petifor, Solid State Commun. 51, 31 (1984). Y. Saad, D. Gao, T. Ngo, S. Bobbit, J. R. Chelikowsky, and W. Andreoni, Phys. Rev. B 85, 104104 (2012).

Proof of Concept: Descriptor for the Classification “Zincblende/Wurtzite (ZB/W) or Rocksalt (RS)?”

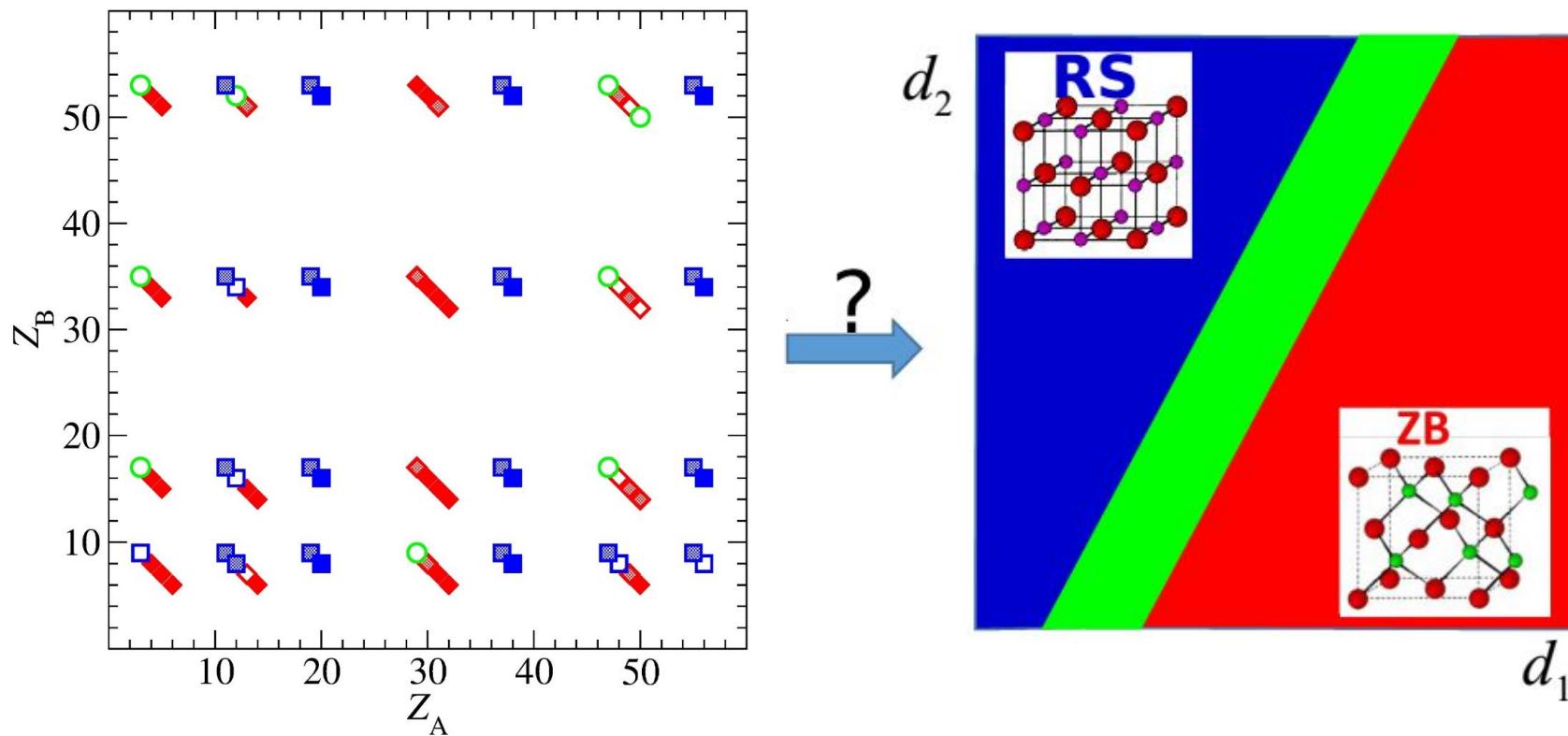
Can we predict not yet calculated structures from Z_A and Z_B ? Can we create a map: “The *ZB/W* community lives here and the *RS* community there?”



- $\Delta = E(\text{RS}) - E(\text{ZB})$
- ◆ ZB, $\Delta > 0.2 \text{ eV}$
- ◆ ZB, $0.1 \text{ eV} < \Delta \leq 0.2 \text{ eV}$
- ◆ ZB, $0.05 \text{ eV} < \Delta \leq 0.1 \text{ eV}$
- $-0.05 \text{ eV} < \Delta \leq 0.05 \text{ eV}$
- RS, $-0.1 \text{ eV} < \Delta \leq -0.05 \text{ eV}$
- RS, $-0.2 \text{ eV} < \Delta \leq -0.1 \text{ eV}$
- RS, $\Delta \leq -0.2 \text{ eV}$

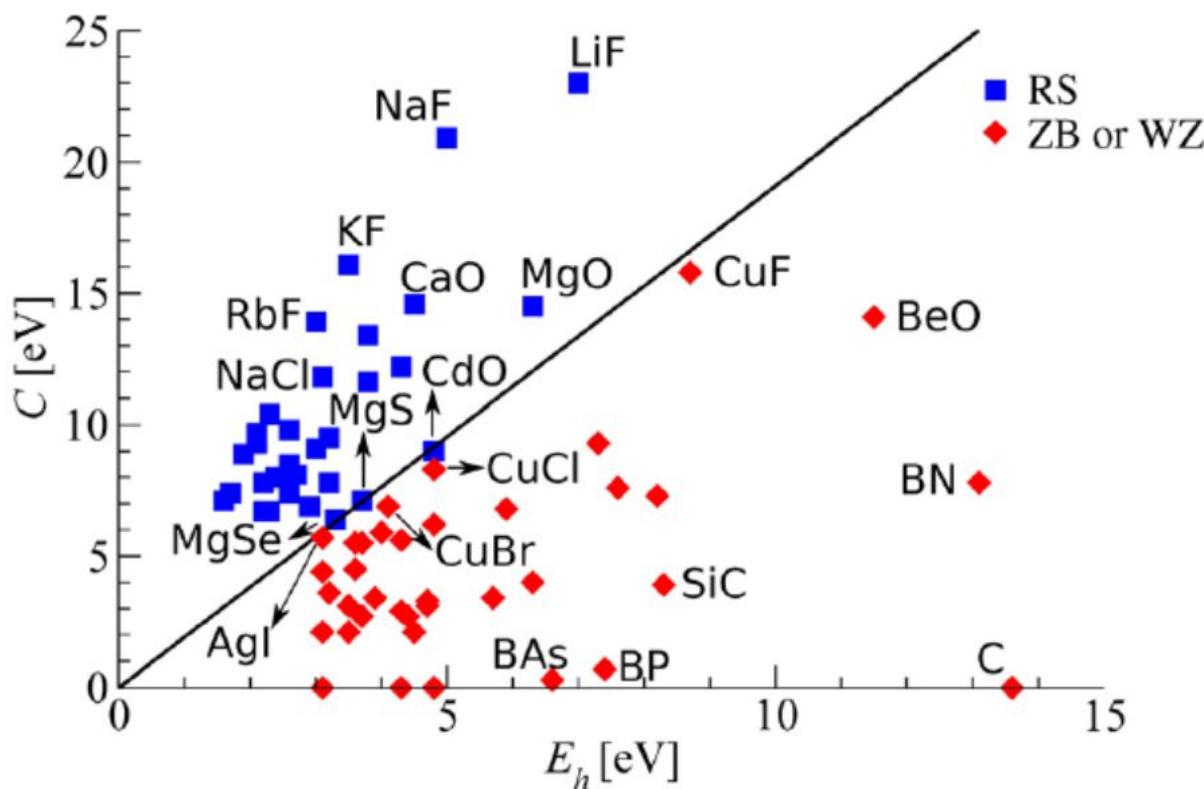
Proof of Concept: Descriptor for the Classification “Zincblende/Wurtzite (ZB/W) or Rocksalt (RS)?”

Can we predict not yet calculated structures from Z_A and Z_B ? Can we create a map: “The ZB/W community has the RS community? No complexity reduction → need a better basis



Proof of Concept: Descriptor for the Classification “Zincblende/Wurtzite (ZB/W) or Rocksalt (RS)?”

Can we predict not yet calculated structures from Z_A and Z_B ? Can we create a map: “The ZB/W community lives here and the RS community there?”



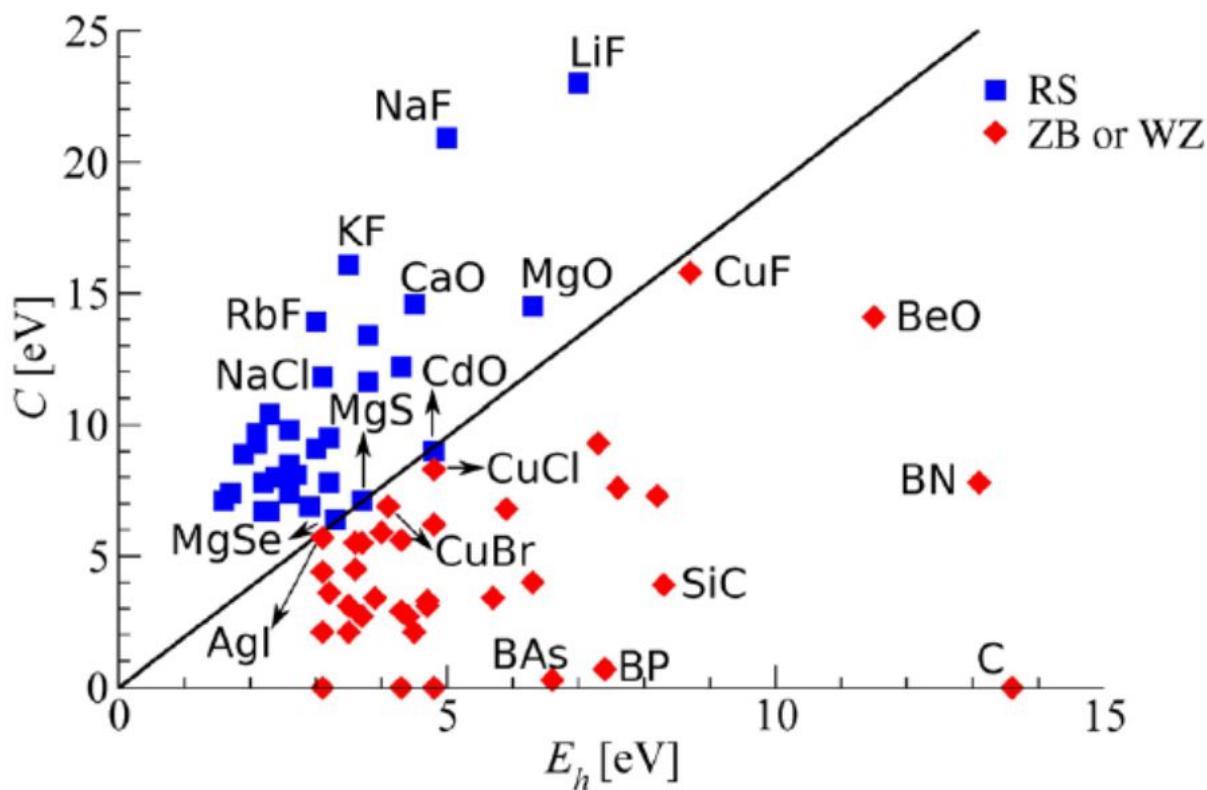
descriptor can be determined
spectroscopically - properties of the solid

J. A. van Vechten, Phys.
Rev. 182, 891 (1969). J. C. Phillips, Rev.
Mod. Phys. 42, 317 (1970).

J. John and A.N. Bloch, Phys. Rev. Lett. 33,
1095 (1974) J. R. Chelikowsky and J. C.
Phillips, Phys. Rev. B 33, 2453 (1978)
A. Zunger, Phys. Rev. B 22, 5839 (1980).
D. G. Petifor, Solid State Commun. 51, 31
(1984). Y. Saad, D. Gao, T. Ngo, S.
Bobbit, J. R. Chelikowsky, and W.
Andreoni, Phys. Rev. B 85, 104104 (2012).

Proof of Concept: Descriptor for the Classification “Zincblende/Wurtzite (ZB/W) or Rocksalt (RS)?”

Can we predict not yet calculated structures from Z_A and Z_B ? Can we create a map: “The ZB/W community lives here and the RS community there?”



descriptor can be determined spectroscopically - properties of the solid

Can we create a map based on calculations simpler than bulk?

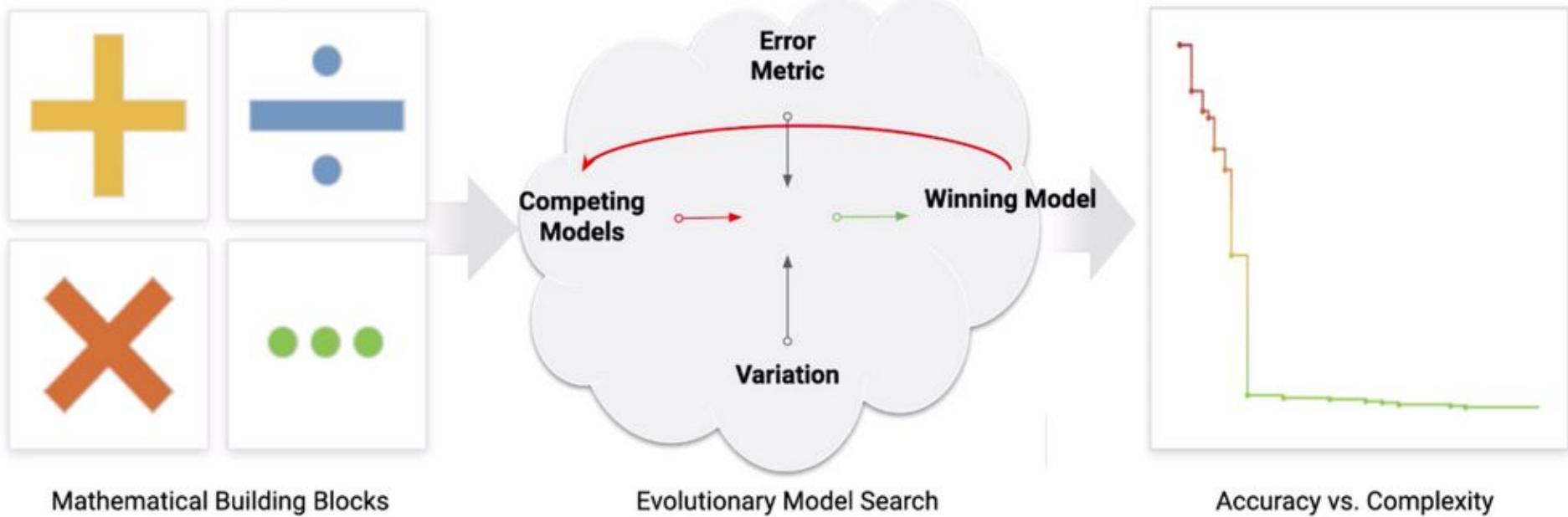
Primary features and feature space

| ID | Description | free atoms | Symbols | # |
|----|---|------------|--|---|
| A1 | Ionization Potential (IP) and Electron Affinity (EA) | | IP(A) EA(A) IP(B) EA(B) [1] | 4 |
| A2 | Highest occupied (H) and lowest unoccupied (L) Kohn-Sham levels | | H(A) L(A) H(B) L(B) | 4 |
| A3 | Radius at the max. value of <i>s</i> , <i>p</i> , and <i>d</i> valence radial radial probability density | | $r_s(A)$ $r_p(A)$ $r_d(A)$ $r_s(B)$ $r_p(B)$ $r_d(B)$ | 6 |

| ID | Description | free dimers | Symbols | # |
|----|----------------------|-------------|-------------------------------|---|
| A4 | Binding energy | | $E_b(AA)$ $E_b(BB)$ $E_b(AB)$ | 3 |
| A5 | HOMO-LUMO KS gap | | HL(AA) HL(BB) HL(AB) | 3 |
| A6 | Equilibrium distance | | $d(AA)$ $d(BB)$ $d(AB)$ | 3 |

How to find the best model for our target property (energy difference between different crystal structures)?

Symbolic regression: Eureqa



Uses evolutionary algorithm to find the best formula describing target property

Assumes “gene” structure of the formula → bias

May result in an unnecessarily complex model

Primary features and feature space

| ID | Description | free atoms | Symbols | # |
|----|---|------------|--|---|
| A1 | Ionization Potential (IP) and Electron Affinity (EA) | | IP(A) EA(A) IP(B) EA(B) [1] | 4 |
| A2 | Highest occupied (H) and lowest unoccupied (L) Kohn-Sham levels | | H(A) L(A) H(B) L(B) | 4 |
| A3 | Radius at the max. value of <i>s</i> , <i>p</i> , and <i>d</i> valence radial radial probability density | | $r_s(A)$ $r_p(A)$ $r_d(A)$ $r_s(B)$ $r_p(B)$ $r_d(B)$ | 6 |

| ID | Description | free dimers | Symbols | # |
|----|----------------------|-------------|-------------------------------|---|
| A4 | Binding energy | | $E_b(AA)$ $E_b(BB)$ $E_b(AB)$ | 3 |
| A5 | HOMO-LUMO KS gap | | HL(AA) HL(BB) HL(AB) | 3 |
| A6 | Equilibrium distance | | $d(AA)$ $d(BB)$ $d(AB)$ | 3 |

| ID | description | prototype formula | # |
|----|---|---|----|
| B1 | absolute differences and sums of A1 | $ IP(A) \pm IP(B) $ | 12 |
| B2 | absolute differences and sums of A2 | $ L(B) \pm H(A) $ | 12 |
| B3 | absolute differences and sums of A3 | $ r_p(A) \pm r_s(A) $ | 30 |
| C3 | squares of A3 and B3 (only sums) | $r_s(A)^2, (r_p(A) + r_s(A))^2$ | 21 |
| D3 | exponentials of A3 and B3 (only sums) | $\exp(r_s(A)), \exp(r_p(A) \pm r_s(A))$ | 21 |
| E3 | exponentials of squared A3 and B3 (only sums) | $\exp(r_s(A)^2), \exp(r_p(A) \pm r_s(A)^2)$ | 21 |

We start with 23 primary features and build > 10,000 non-linear combinations

Mathematical formulation of the problem

P_j -- property value ($E_{ZB} - E_{RS}$) for material j (a function in materials space)

$d_{j,l}$ -- value of feature l related to material j (e.g., $|r_s(A_j) - r_p(B_j)|$) (a basis function in materials space)

c_l -- coefficient of the expansion of the property function in terms of basis functions:

$$P_j = \sum_l d_{j,l} c_l \quad \text{How to find } c_l?$$

$$\sum_j \left(P_j - \sum_l d_{j,l} c_l \right)^2 + \lambda \|c\|_n \rightarrow \text{argmin}(c)$$

regularization term to explore and ensure compactness of the expansion (reduce complexity)

Mathematical formulation of the problem

P_j -- property value ($E_{ZB} - E_{RS}$) for material j (a function in materials space)

$d_{j,l}$ -- value of feature l related to material j (e.g., $|r_s(A_j) - r_p(B_j)|$) (a basis function in materials space)

c_l -- coefficient of the expansion of the property function in terms of basis functions:

$$P_j = \sum_l d_{j,l} c_l \quad \text{How to find } c_l?$$

$$\sum_j \left(P_j - \sum_l d_{j,l} c_l \right)^2 + \lambda \|c\|_n \rightarrow \operatorname{argmin}(c)$$

$\|c\|_0$ -- number of non-zero coefficients \rightarrow NP hard! (need to try all combinations)

Mathematical formulation of the problem

P_j -- property value ($E_{ZB} - E_{RS}$) for material j (a function in materials space)

$d_{j,l}$ -- value of feature l related to material j (e.g., $|r_s(A_j) - r_p(B_j)|$) (a basis function in materials space)

c_l -- coefficient of the expansion of the property function in terms of basis functions:

$$P_j = \sum_l d_{j,l} c_l \quad \text{How to find } c_l?$$

$$\sum_j \left(P_j - \sum_l d_{j,l} c_l \right)^2 + \lambda \|c\|_n \rightarrow \operatorname{argmin}(c)$$

$\|c\|_0$ -- number of non-zero coefficients \rightarrow NP hard! (need to try all combinations)

$\|c\|_2 = \sum_l |c_l|^2$ -- ridge regression \rightarrow not most compact!

$\|c\|_1 = \sum_l |c_l|$ -- LASSO (Least Absolute Shrinkage and Selection Operator) \rightarrow convex problem, equivalent to the NP-hard if features (columns of d) are uncorrelated

Compressed (compressive?) sensing



Raw: 15MB



JPEG: 150KB

Expand in a basis (wavelets) → use LASSO to select most important basis functions → store compressed image

Mathematical formulation of the problem

P_j -- property value ($E_{ZB} - E_{RS}$) for material j (a function in materials space)

$d_{j,l}$ -- value of feature l related to material j (e.g., $|r_s(A_j) - r_p(B_j)|$) (a basis function in materials space)

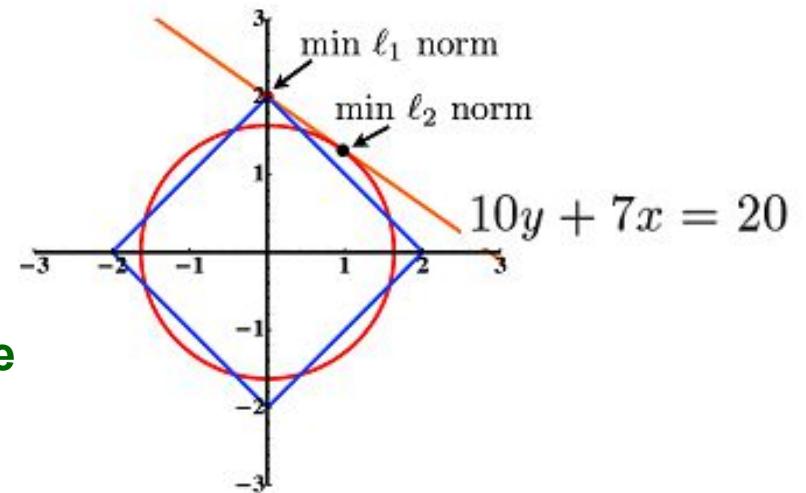
c_l -- coefficient of the expansion of the property function in terms of basis functions:

$$P_j = \sum_l d_{j,l} c_l$$

How to find c_l ?

$$\sum_j \left(P_j - \sum_l d_{j,l} c_l \right)^2 + \lambda \|c\|_n \rightarrow \text{argmin}(c)$$

$\|c\|_1 = \sum_l |c_l|$ -- LASSO (Least Absolute Shrinkage and Selection Operator) \rightarrow convex problem, equivalent to the NP-hard if features (columns of D) are uncorrelated (no linear dependence in the basis set)



The descriptors selected with LASSO

$$\frac{IP(B) - EA(B)}{r_p(A)^2} \underset{1D}{\boxed{\text{1D}}}, \frac{|r_s(A) - r_p(B)|}{\exp(r_s(A))} \underset{2D}{\boxed{\text{2D}}}, \frac{|r_p(B) - r_s(B)|}{\exp(r_d(A))} \underset{3D}{\boxed{\text{3D}}}$$

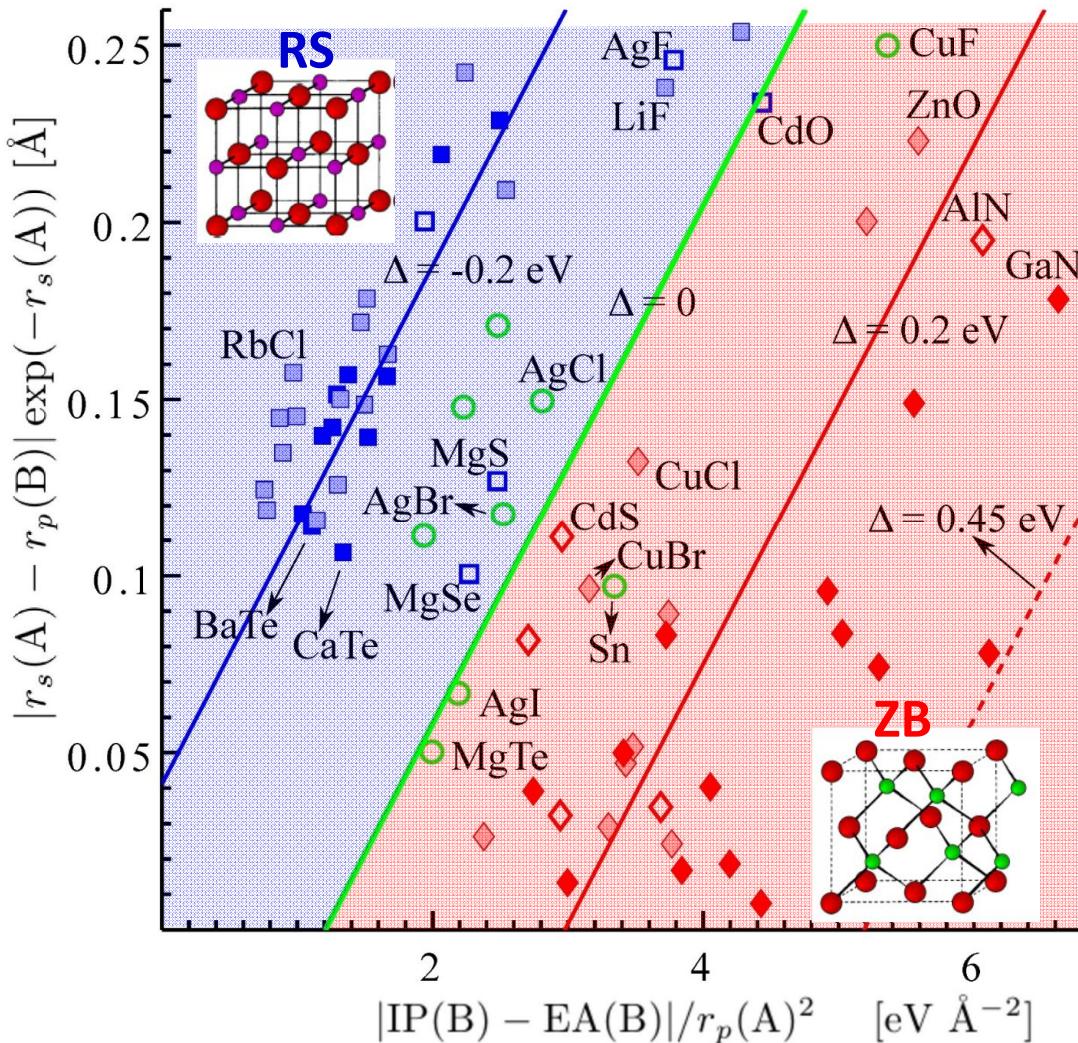
$$\Delta E = 0.117 \frac{EA(B) - IP(B)}{r_p(A)^2} - 0.342 \quad \text{1D}$$

$$\Delta E = 0.113 \frac{EA(B) - IP(B)}{r_p(A)^2} + 1.542 \frac{|r_s(A) - r_p(B)|}{\exp(r_s(A))} - 0.137 \quad \text{2D}$$

$$\begin{aligned} \Delta E = & 0.108 \frac{EA(B) - IP(B)}{r_p(A)^2} + 1.790 \frac{|r_s(A) - r_p(B)|}{\exp(r_s(A))} + \\ & + 3.766 \frac{|r_p(B) - r_s(B)|}{\exp(r_d(A))} - 0.0267 \end{aligned} \quad \text{3D}$$

Same features are selected for higher-dimensional descriptors, but this does not have to be the case

“The Map” -- compressed sensing -- LASSO, 2D descriptor



- $\Delta = E(\text{RS}) - E(\text{ZB})$
- ◆ ZB, $\Delta > 0.2 \text{ eV}$
- ◆ ZB, $0.1 \text{ eV} < \Delta \leq 0.2 \text{ eV}$
- ◆ ZB, $0.05 \text{ eV} < \Delta \leq 0.1 \text{ eV}$
- $-0.05 \text{ eV} < \Delta \leq 0.05 \text{ eV}$
- RS, $-0.1 \text{ eV} < \Delta \leq -0.05 \text{ eV}$
- RS, $-0.2 \text{ eV} < \Delta \leq -0.1 \text{ eV}$
- RS, $\Delta \leq -0.2 \text{ eV}$

$$P(j) = \mathbf{d}(j)\mathbf{c}$$

The complexity and science is
in the descriptor (identified
from >10,000 features).

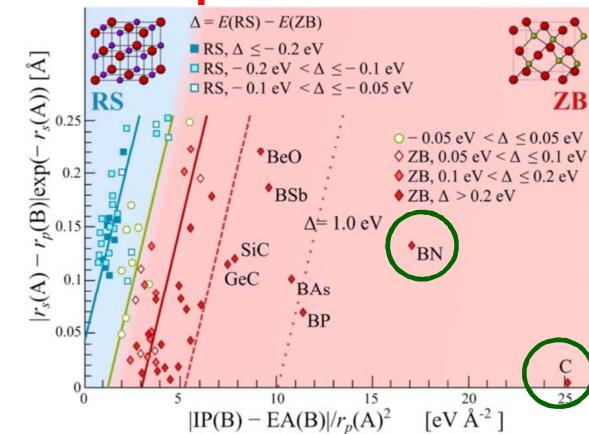
L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko,
C. Draxl, and M. Scheffler,
Phys. Rev. Lett. **114**, 105503 (2015).

Predictive power of the model

Hadn't we known about diamond ... we'd have predicted it!

When both carbon diamond and BN are excluded from training:

| | $\Delta E(\text{LDA})$ | $\Delta E(\text{predicted})$ |
|----|------------------------|------------------------------|
| C | -2.64 eV | -1.44 eV |
| BN | -1.71 eV | -1.37 eV |



Hadn't we known about any carbon-containing binary ... we'd have predicted carbon chemistry (from atomic features)

If all C containing binaries (C, SiC, GeC, and SnC) are excluded from training, i.e. no explicit information on C is given to the model:

| | $\Delta E(\text{LDA})$ | $\Delta E(\text{predicted})$ |
|-----|------------------------|------------------------------|
| C | -2.64 eV | -1.37 eV |
| SiC | -0.67 eV | -0.48 eV |
| GeC | -0.81 eV | -0.46 eV |
| SnC | -0.45 eV | -0.23 eV |

Predictive power of the model

Mean absolute error (MAE), and maximum absolute error (MaxAE), in eV, (first two lines) and for a leave-10%-out cross validation (CV), averaged over 150 random selections of the training set (last two lines). For (Z_A^*, Z_B^*) , each atom is identified by a string of three random numbers.

| Descriptor | Z_A, Z_B | Z_A^*, Z_B^* | 1D | 2D | 3D | 5D |
|------------|-------------------|-------------------|------|------|------|------|
| MAE | $1 \cdot 10^{-4}$ | $3 \cdot 10^{-3}$ | 0.12 | 0.08 | 0.07 | 0.05 |
| MaxAE | $8 \cdot 10^{-4}$ | 0.03 | 0.32 | 0.32 | 0.24 | 0.20 |
| MAE, CV | 0.13 | 0.14 | 0.12 | 0.09 | 0.07 | 0.05 |
| MaxAE, CV | 0.43 | 0.42 | 0.27 | 0.18 | 0.16 | 0.12 |

Gaussian-kernel ridge regression

LASSO

Predictive power of the model

Mean absolute error (MAE), and maximum absolute error (MaxAE), in eV, (first two lines) and for a leave-10%-out cross validation (CV), averaged over 150 random selections of the training set (last two lines). **For (Z_A^*, Z_B^*) , each atom is identified by a string of three random numbers.**

| Descriptor | Z_A, Z_B | Z_A^*, Z_B^* | 1D | 2D | 3D | 5D |
|------------|-------------------|-------------------|------|------|------|------|
| MAE | $1 \cdot 10^{-4}$ | $3 \cdot 10^{-3}$ | 0.12 | 0.08 | 0.07 | 0.05 |
| MaxAE | $8 \cdot 10^{-4}$ | 0.03 | 0.32 | 0.32 | 0.24 | 0.20 |
| MAE, CV | 0.13 | 0.14 | 0.12 | 0.09 | 0.07 | 0.05 |
| MaxAE, CV | 0.43 | 0.42 | 0.27 | 0.18 | 0.16 | 0.12 |

Gaussian-kernel ridge regression **LASSO**

Predictive power of the model

Mean absolute error (MAE), and maximum absolute error (MaxAE), in eV, (first two lines) and for a leave-10%-out cross validation (CV), averaged over 150 random selections of the training set (last two lines). For (Z_A^*, Z_B^*) , each atom is identified by a string of three random numbers.

| Descriptor | Z_A, Z_B | Z_A^*, Z_B^* | 1D | 2D | 3D | 5D |
|------------|-------------------|-------------------|------|------|------|------|
| MAE | $1 \cdot 10^{-4}$ | $3 \cdot 10^{-3}$ | 0.12 | 0.08 | 0.07 | 0.05 |
| MaxAE | $8 \cdot 10^{-4}$ | 0.03 | 0.32 | 0.32 | 0.24 | 0.20 |
| MAE, CV | 0.13 | 0.14 | 0.12 | 0.09 | 0.07 | 0.05 |
| MaxAE, CV | 0.43 | 0.42 | 0.27 | 0.18 | 0.16 | 0.12 |

Gaussian-kernel ridge regression

LASSO

Drawing causal inference from data



a mapping exists, even a physical intuition exists, but ΔE does not listen directly to the descriptor (intricate causality)

$$P(j) = d(j)c$$

There are two aspects:

- 1) practical aspect -- we benefit from knowing $d \rightarrow P$ mapping for any convenient $d(j)$ (analogy: plane waves)
- 2) physical aspect (understanding) -- we can reduce the complexity of the model and at the same time increase its applicability domain by a clever choice of $d(j)$ (analogy: atomic orbitals and molecular-orbital picture)

We greatly benefit from $d(j)$ providing a framework for a rational analysis

CH_4 chemical decomposition under shock-compression conditions (high T and p)

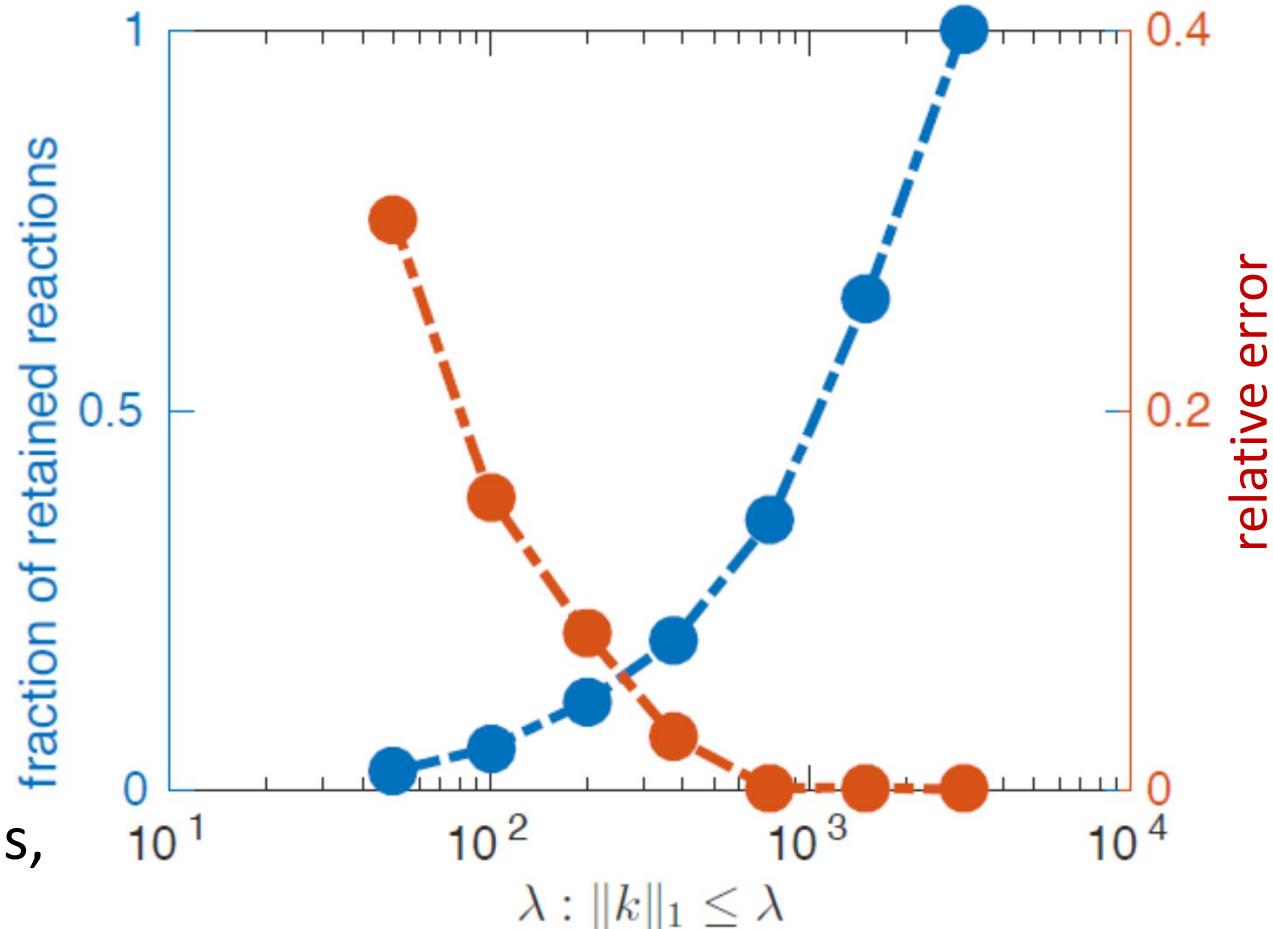
Yang, Q., Sing-Long, C. A., Reed, E. J., MRS Advances 1 (2016)

**Methane at $T = 3,300 \text{ K}$,
 $p = 40.53 \text{ GPa}$:** MD simulations (using a force-field description) find 2,613 different chemical reactions. Using compressed sensing it is shown that only 11% of them are relevant.

$$\min_{\hat{\mathbf{k}}} \|A\hat{\mathbf{k}} - \mathbf{b}\|_2$$

subject to $\hat{\mathbf{k}} \geq 0, \|\hat{\mathbf{k}}\|_1 \leq \lambda$

The A matrix has 2,613 columns, 2,395,918,510 rows



Lattice Anharmonicity and Thermal Conductivity from Compressive Sensing of First-Principles Calculations

$$F_a = -\Phi_a - \Phi_{ab} u_b - \frac{1}{2} \Phi_{abc} u_b u_c - \dots$$

force on atom a
(training data)
force constant tensor
 $\partial^2 E / \partial u_a \partial u_b$
(unknown)
displacement of atom c
(training data)

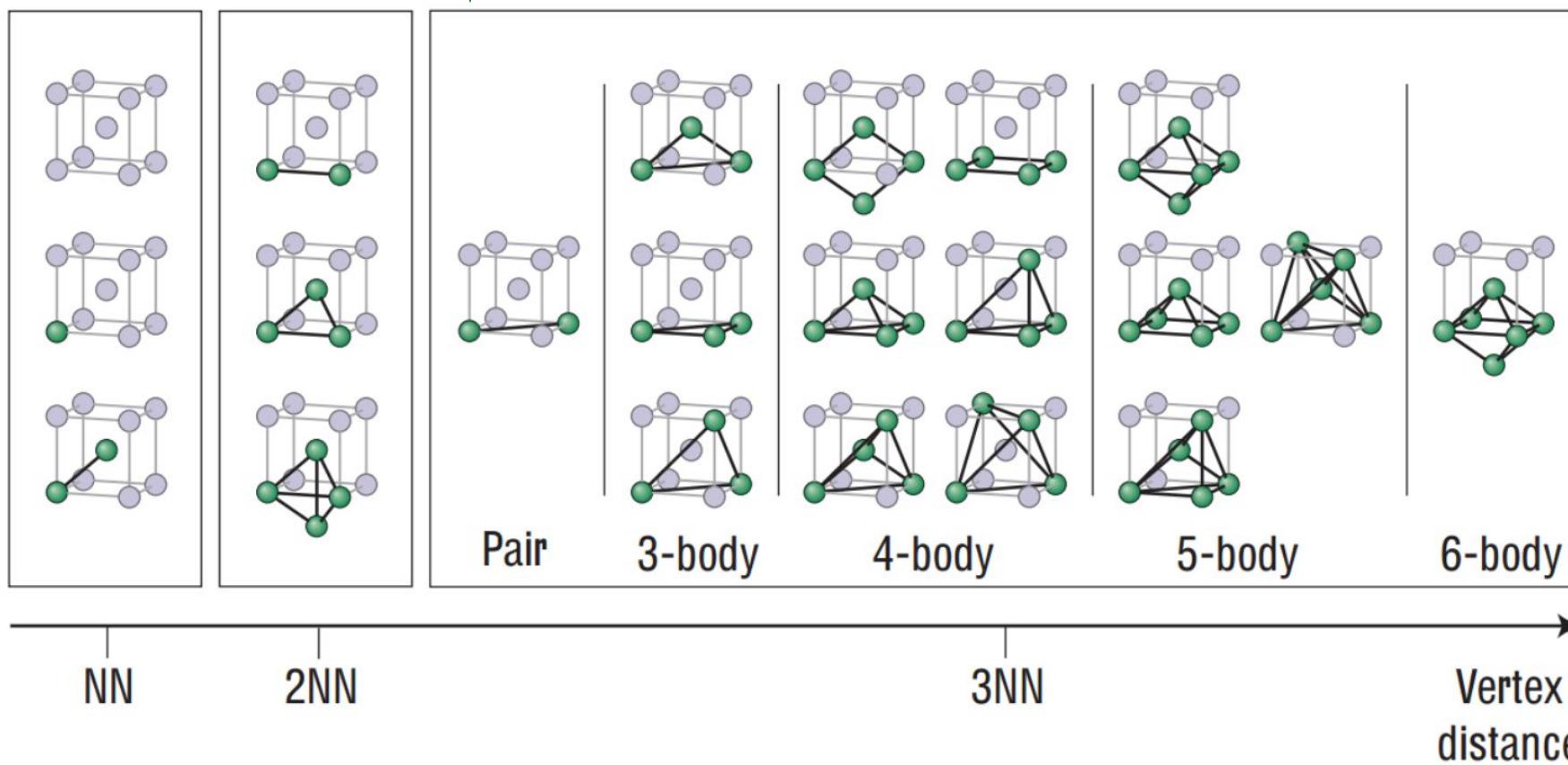
$$\min_{\Phi} \left(\lambda \sum_I |\Phi_I| + \sum_a (F_a - A_{aJ} \Phi_J)^2 \right) \rightarrow \Phi$$

$$A_{aJ} = \begin{bmatrix} -1 & u_b^1 & -\frac{1}{2} u_b^1 u_c^1 & \dots \\ \vdots & \vdots & \vdots & \ddots \\ -1 & u_b^L & -\frac{1}{2} u_b^L u_c^L & \dots \end{bmatrix}$$

→ predictive model for anharmonic lattice dynamics

Compressive Sensing for Cluster Expansion

$$E(\sigma) = E_0 + \sum_f \Pi_f(\sigma) J_f$$
$$\min_{J_f} \left(\lambda \sum_f |J_f| + \sum_i (E^{DFT}(\sigma_i) - E^{CE}(\sigma_i))^2 \right) \rightarrow J_f$$



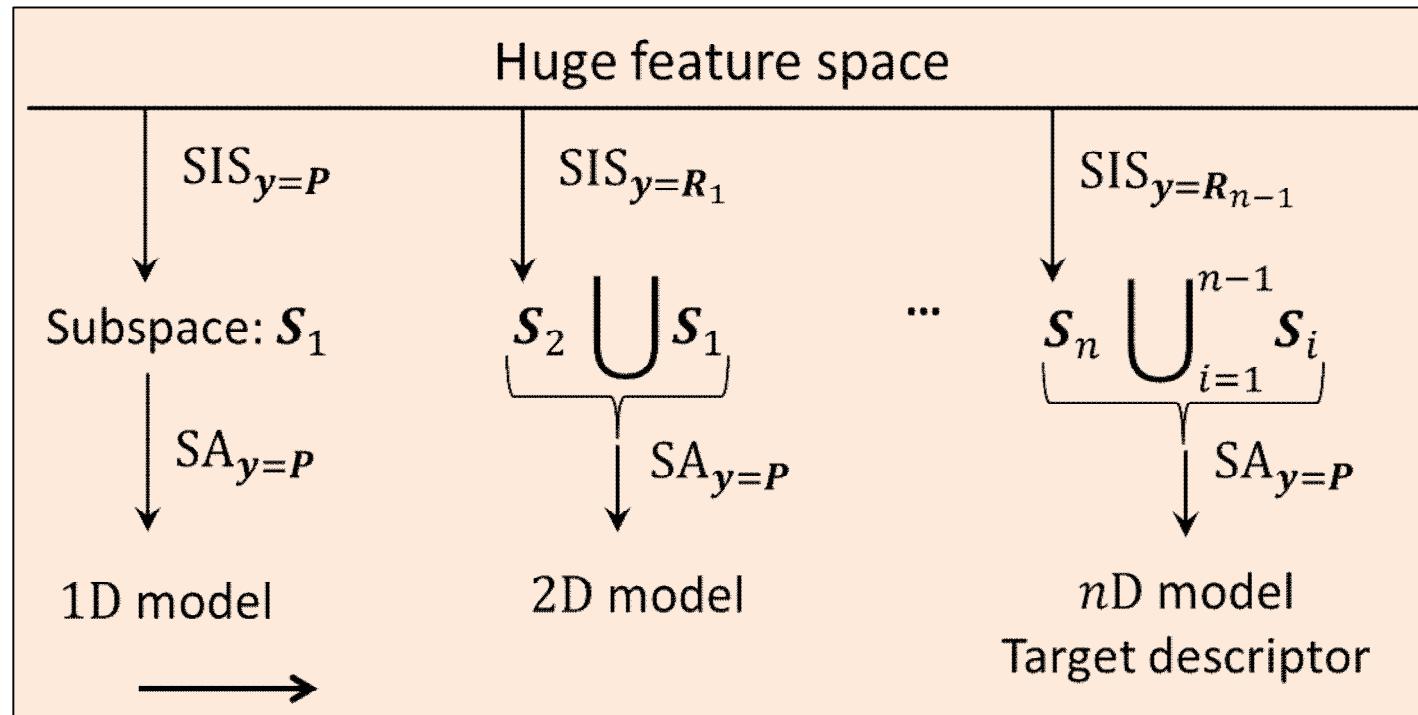
L. J. Nelson, G. L. W. Hart, F. Zhou, and V. Ozoliņš, Phys. Rev. B 87, 035125 (2013)

Enabling Feature Spaces with Billions of Elements by Sure Independence Screening

$\|c\|_1 = \sum_l |c_l|$ -- LASSO \rightarrow convex problem, equivalent to the NP-hard if features are uncorrelated \rightarrow not the case when many features are generated \rightarrow Sure Independence Screening plus Selection Operator (SISSO)

1. Systematically construct a huge feature space (10^{11}) from primary features: $\hat{R} = \{+, -, :, ^{-1}, ^2, ^3, \sqrt{}, \exp, \log, |-|\}$ (use physically meaningful combinations!)
2. Select top ranked features using *Sure Independence Screening (SIS)*^[1] (correlation learning). Select n features corresponding to the n largest projection on the target property, i.e. largest components of the vector ($D^T y$)
 - y : vector with the target property (e.g., rock salt-zincblende energy differences; 82 elements)
 - D : matrix of the feature space (e.g., 82 x 100 billion elements)
3. Apply a sparsifying operator (l_0 regularization) to the selected features to determine 1D, 2D,... descriptors

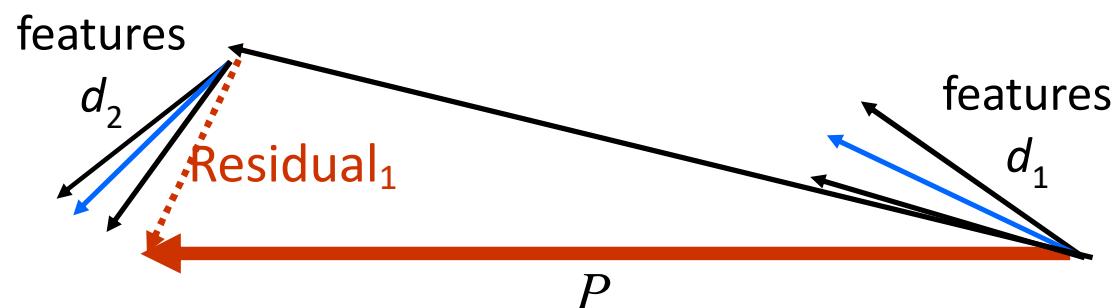
SISSO: Iterative residual fitting



y : response vector

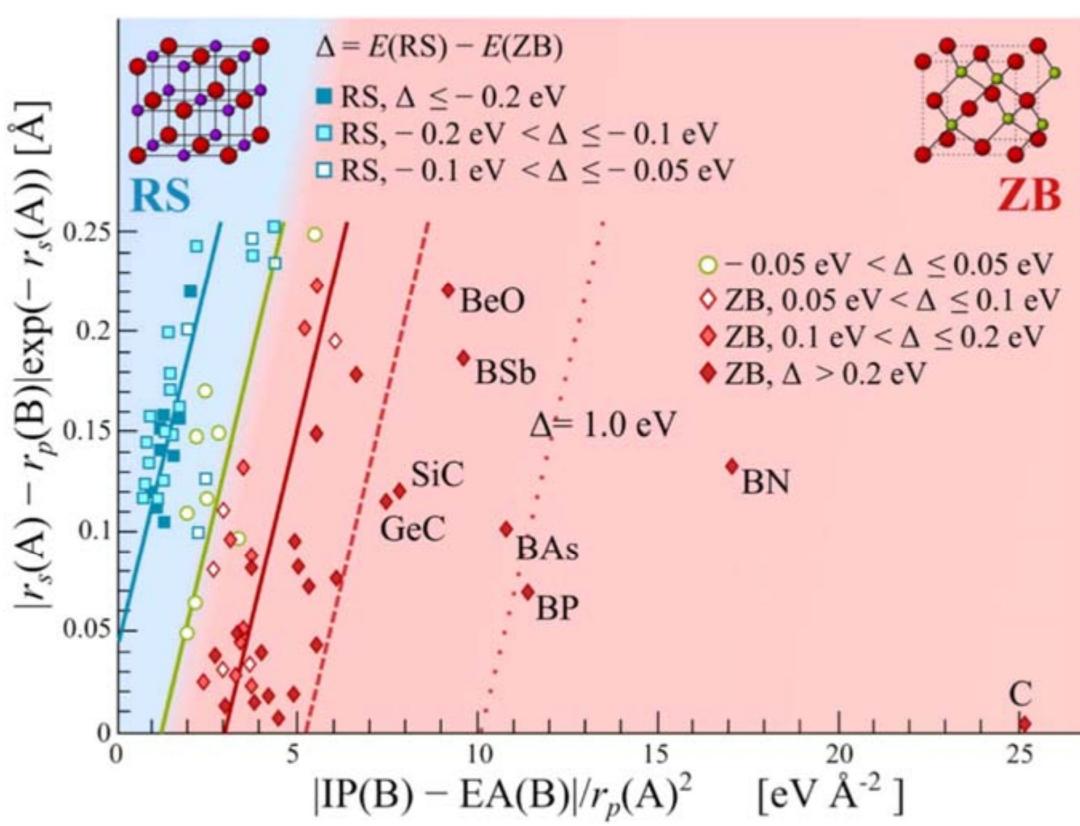
P : target material property

Residual: $R = P - \sum_i c_i d_i$

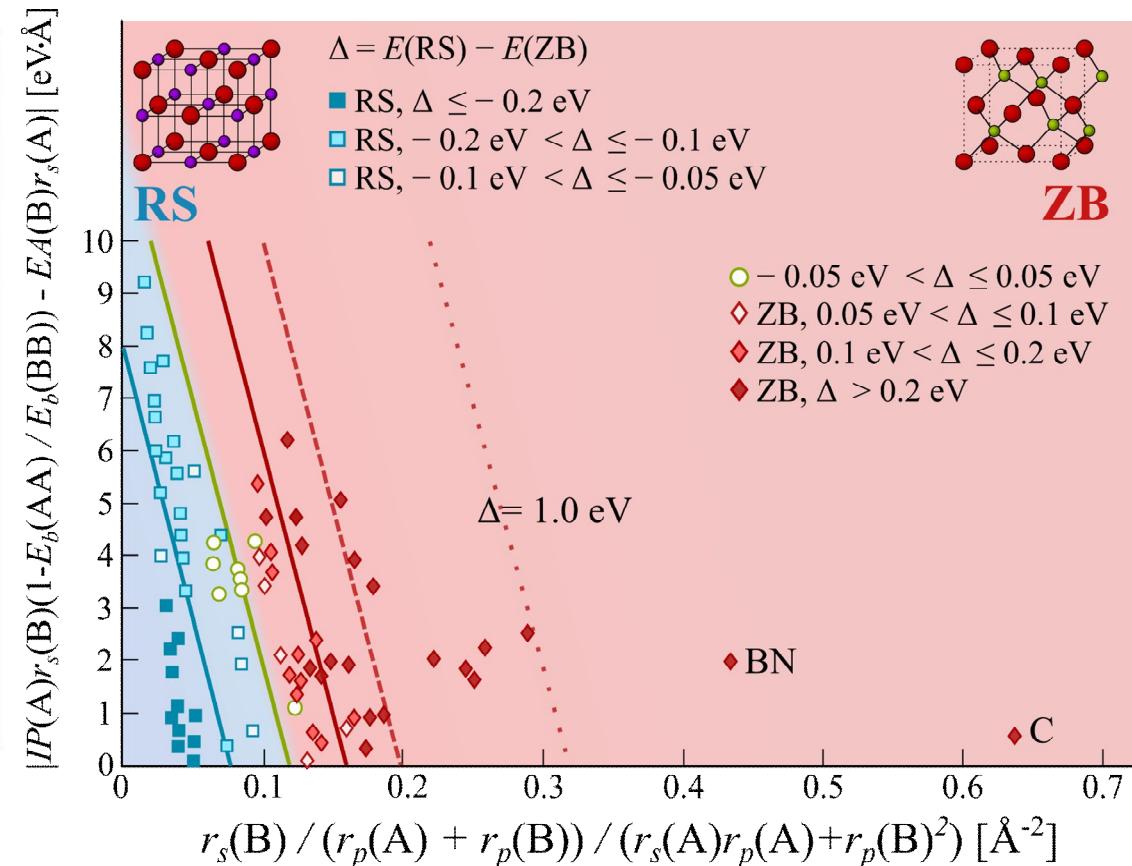


SISSO: Performance

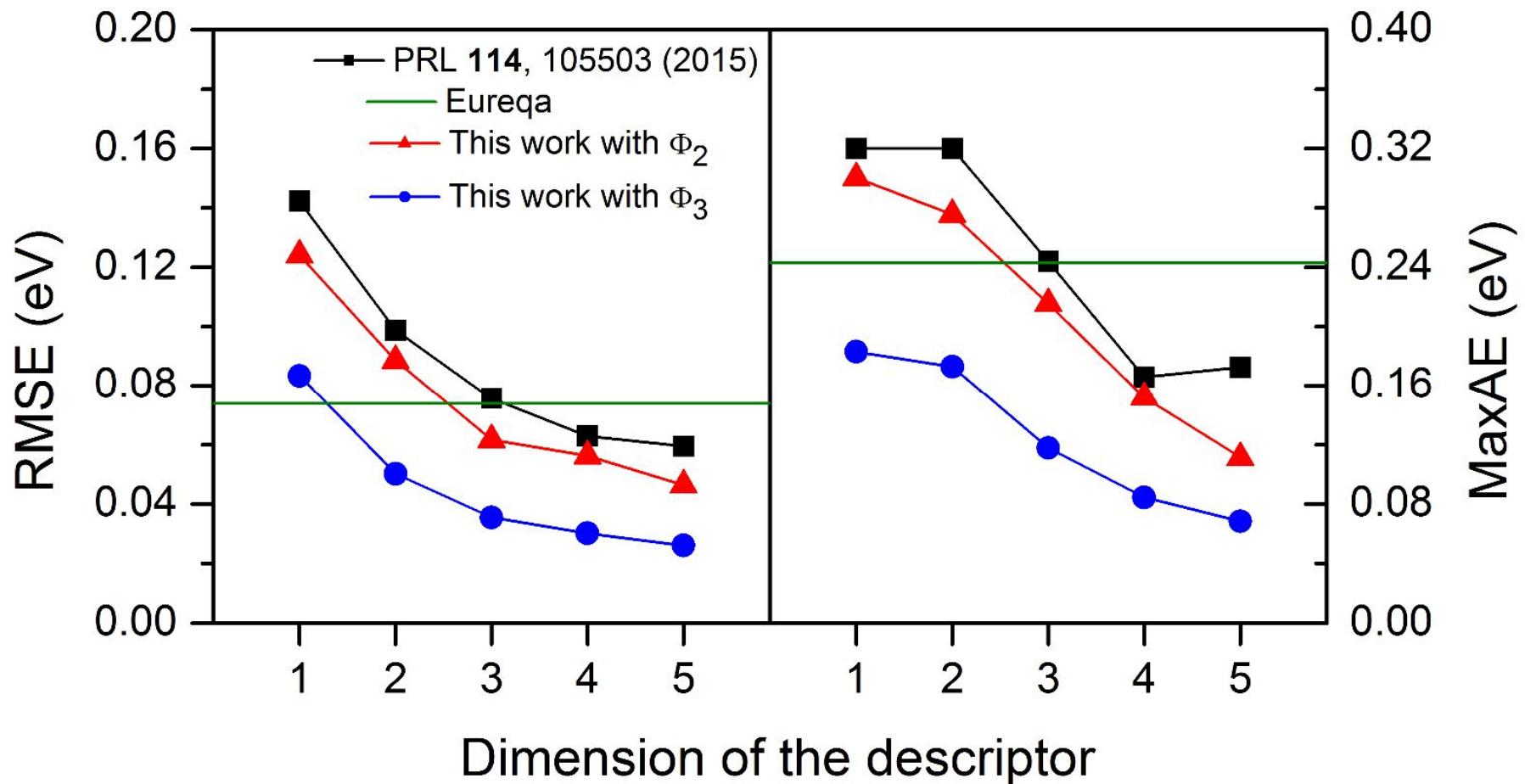
LASSO($+l_0$)



SISSO



SISSO: Performance



SISSO: Multitask and categorical

Multitask: Construct simultaneously SISSO models for several properties with the same descriptor

$$\min_{\mathbf{c}} \left(\lambda \|c_i^k\|_0 + \sum_k \frac{1}{N_{\text{samples}}^k} \sum_{\substack{\text{samples} \\ \text{in } k}} (P^k - \mathbf{d}\mathbf{c}^k)^2 \right) \rightarrow \mathbf{c}$$

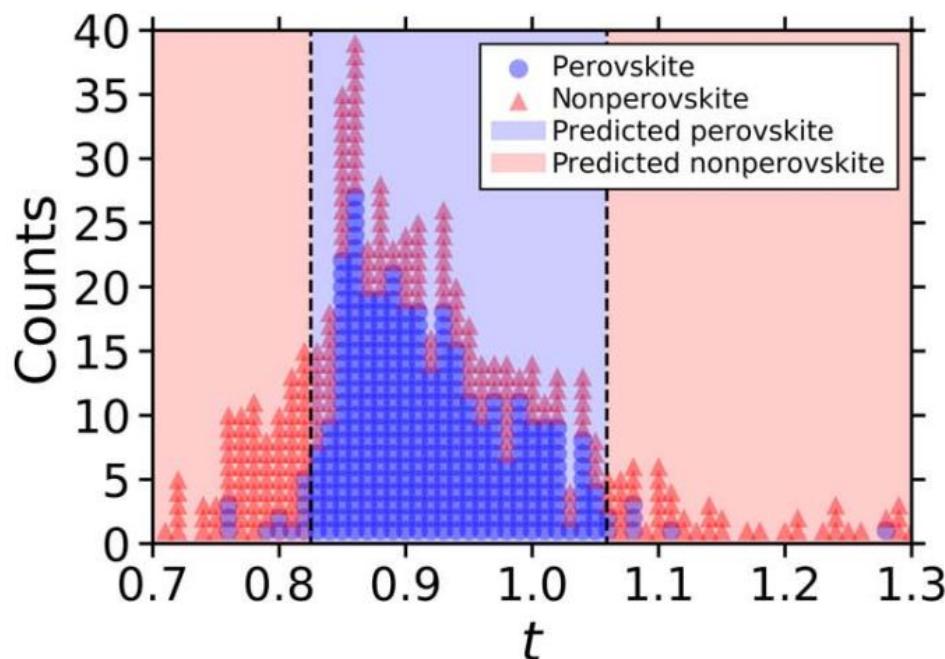
Categorical (can be also multitask): Property - material belongs to a given class (yes/no)

$$\min_{\mathbf{c}} \left(\lambda \|c_i^k\|_0 + \sum_{I=1}^{N_{\text{classes}}} \sum_{J \neq I} O_{IJ}(\mathbf{d}, \mathbf{c}) \right) \rightarrow \mathbf{c}$$

number of data in the overlap region between domains of different classes in d -space

SISSO: Examples

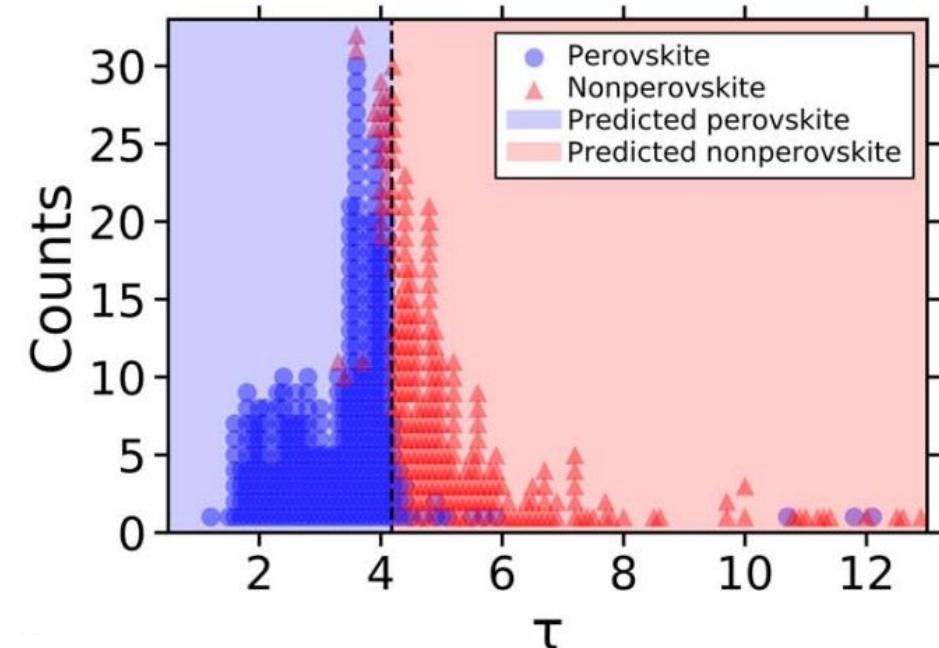
- Perovskite phase stability (improved tolerance factor)



Goldschmidt factor: accuracy 79%

$$0.825 < \frac{r_A+r_X}{\sqrt{2}(r_B+r_X)} < 1.059$$

↗ ionic radii

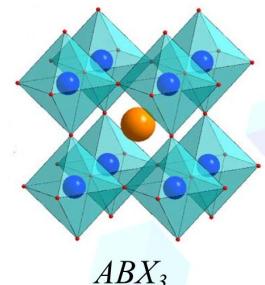


New factor: accuracy 92%

$$\frac{r_X}{r_B} - n_A \left(n_A - \frac{r_A/r_B}{\ln(r_A/r_B)} \right) < 4.18$$

↗ oxidation state

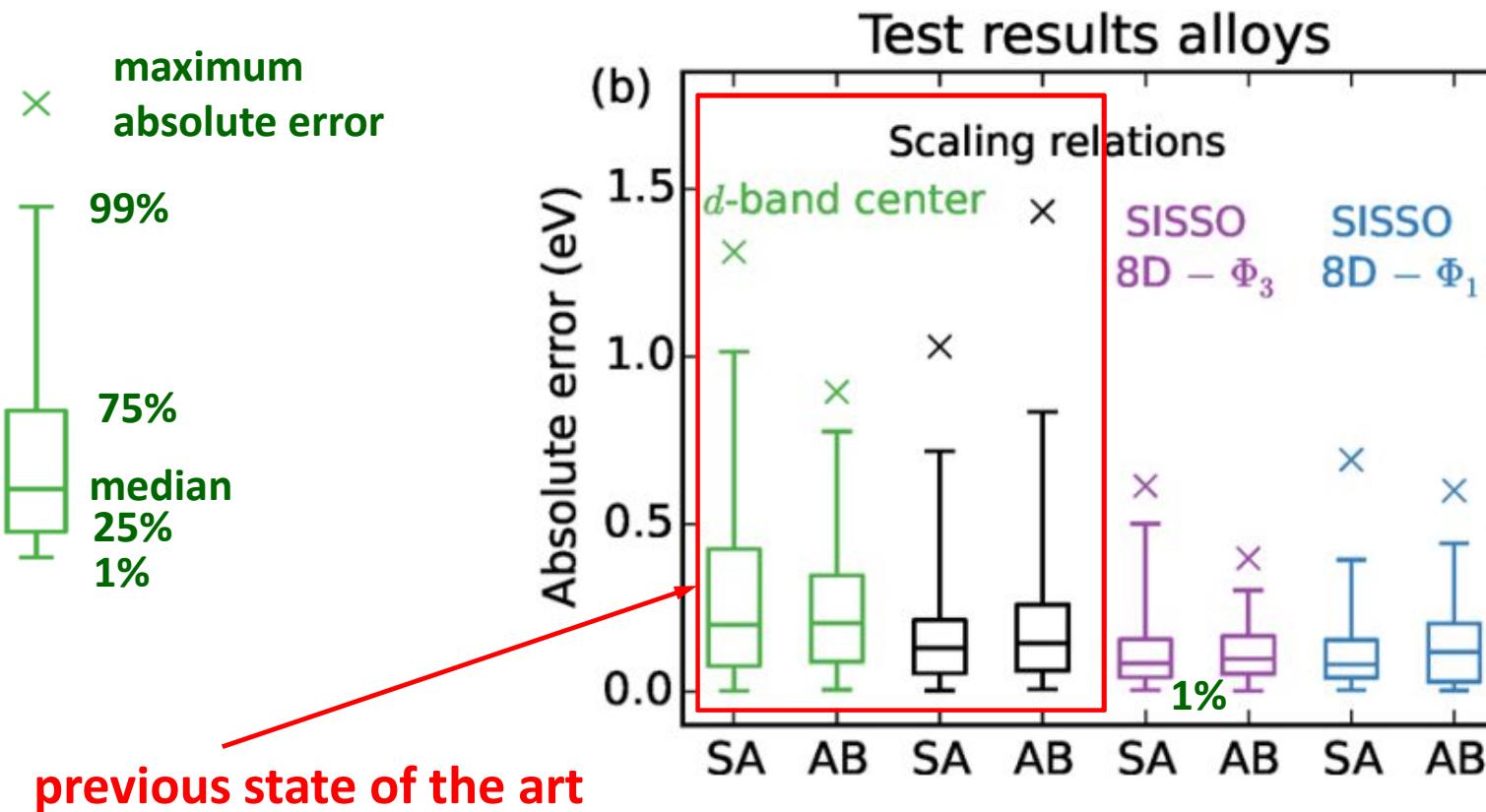
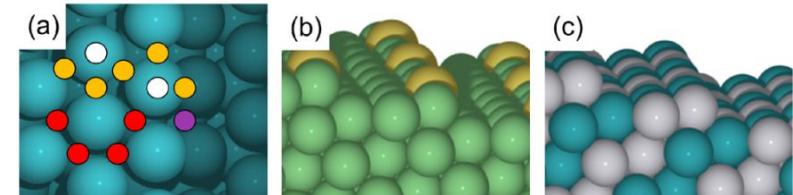
C. Bartel *et al.*, Sci. Adv. 5, eaav0693 (2019)



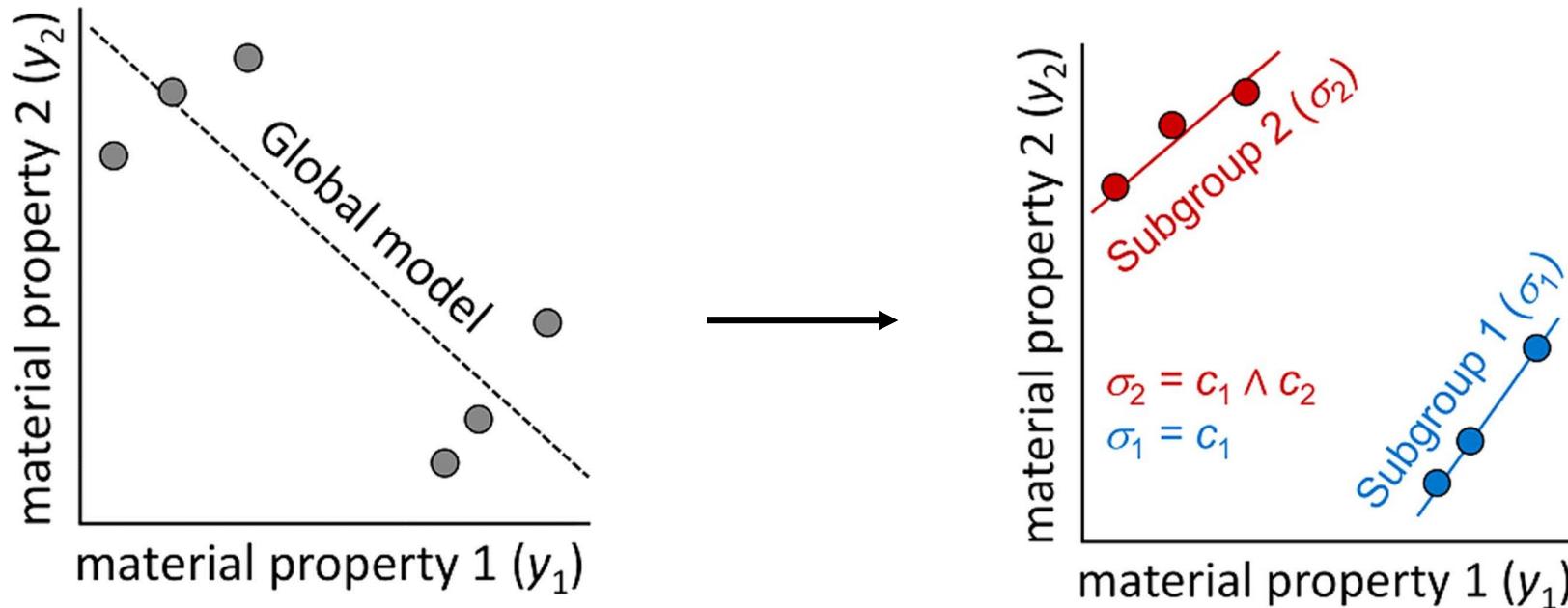
SISSO: Examples

- Adsorption of molecules on metal surfaces

Adsorption of C, CH, CO, H, O, OH)



Data mining: Subgroup discovery



Subgroups are defined by selectors σ expressed as “AND” combinations of statements like “band gap < 2 eV”, “atom radius > 1.4 Å”, etc.

SGD algorithm: find subgroups that maximize *quality function*

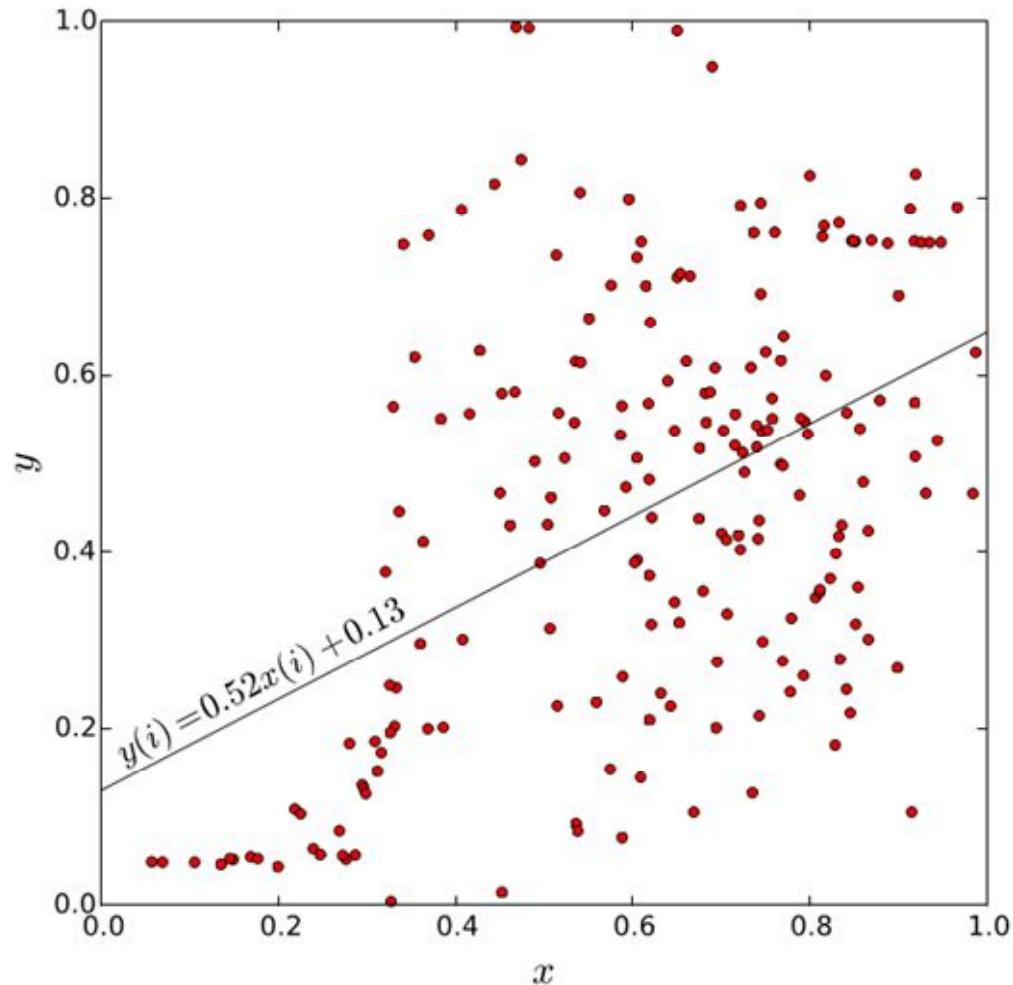
$$f = N_{\text{subgroup}} / N_{\text{all}} \times |\text{mean}_{\text{subgroup}} - \text{mean}_{\text{all}}| \times (1 - \text{variance}_{\text{subgroup}} / \text{variance}_{\text{all}})$$

Numerical separators (“band gap < 2 eV”) from k-means clustering (unsupervised learning)

Search for subgroups: Monte Carlo or branch-and-bound algorithm

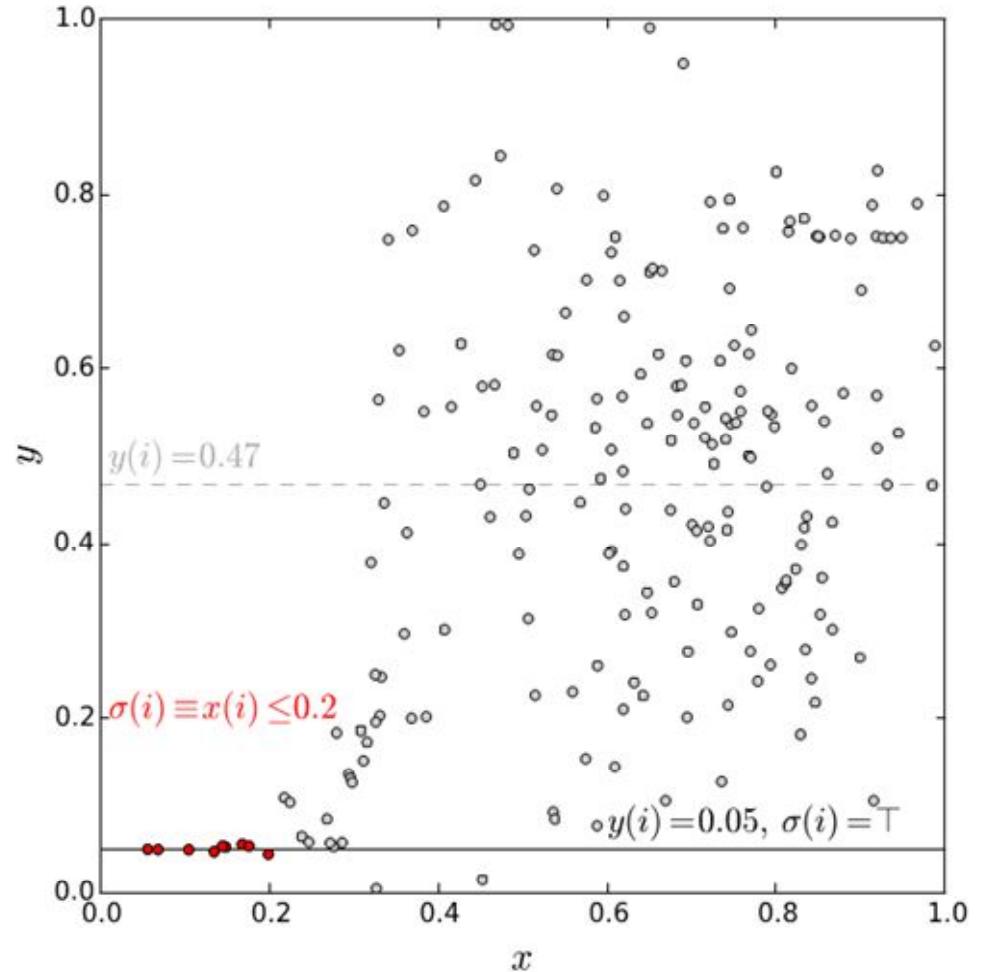
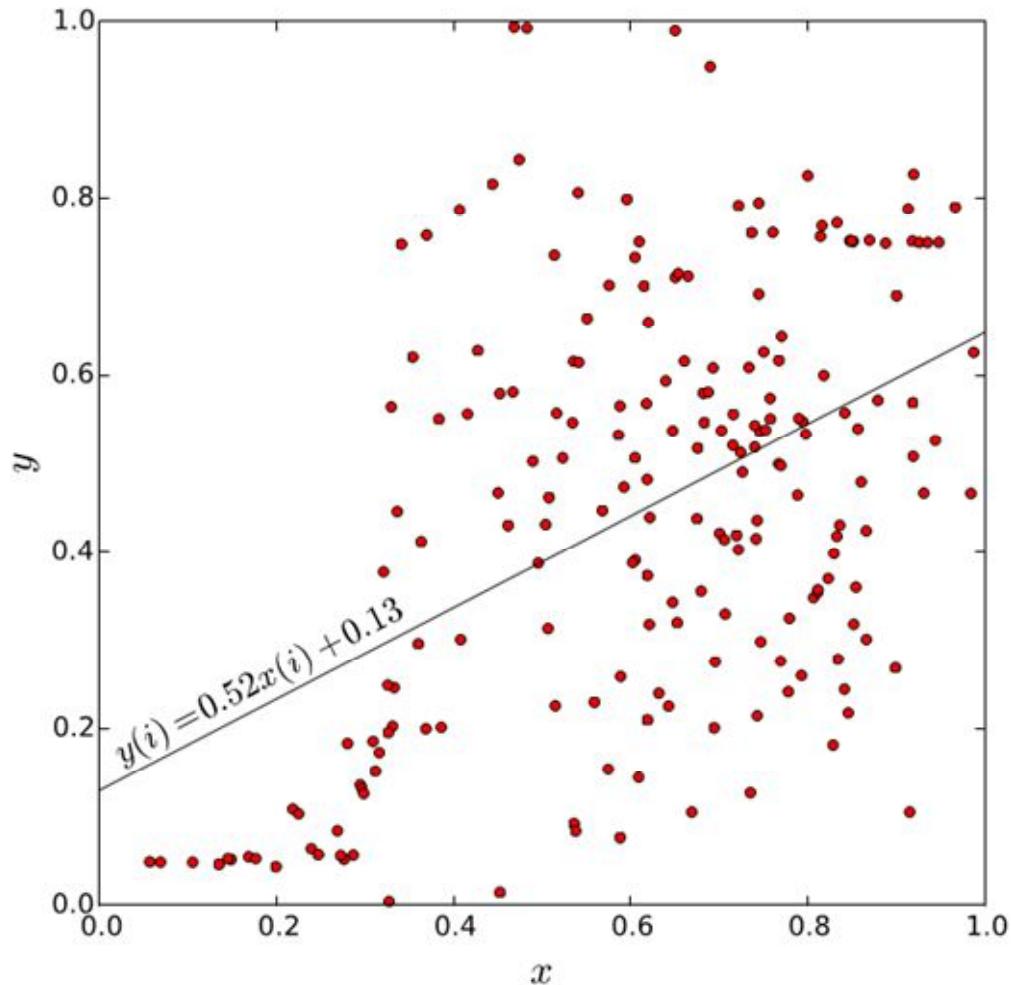
W. Klösgen, Advances in Knowledge Discovery and Data Mining. Palo Alto, CA: AAAI Press; 1996, 249

Data mining: Subgroup discovery

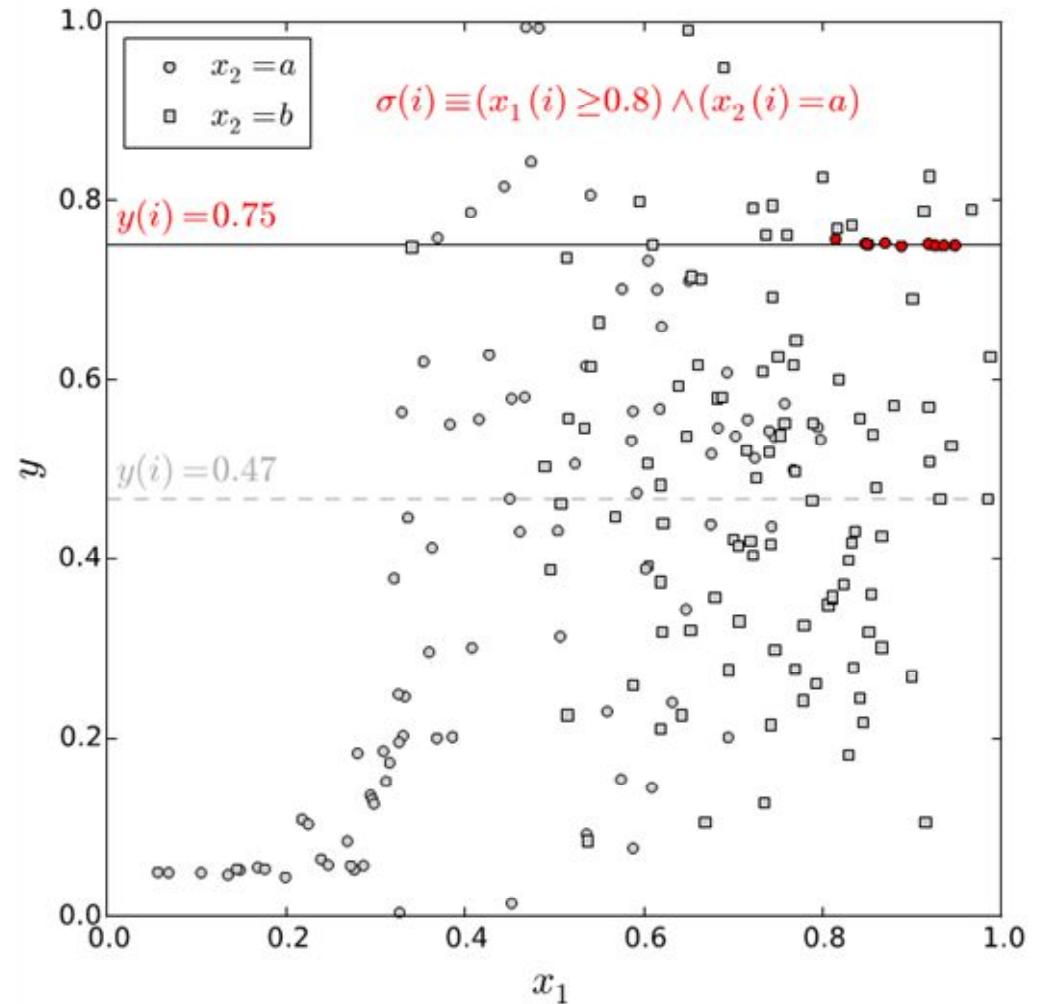
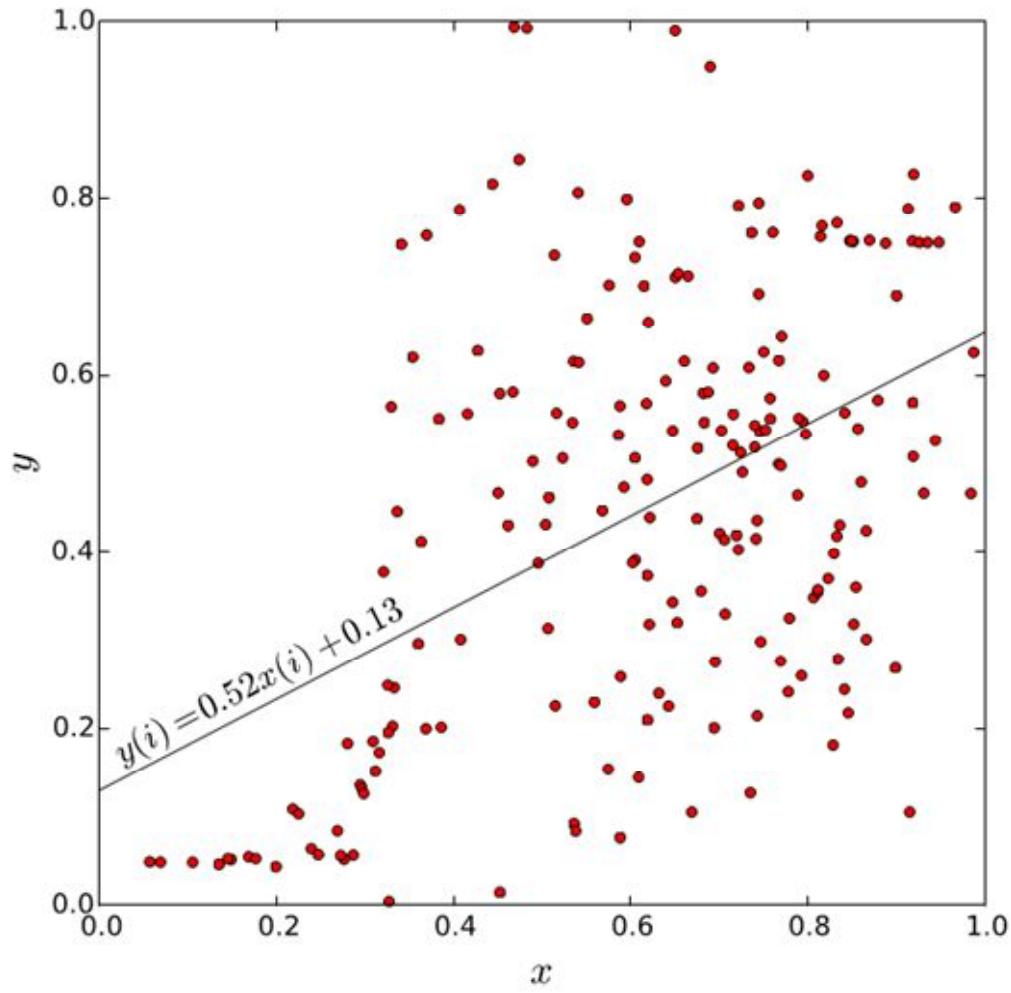


M. Boley *et al.*, Data Min. Knowl. Disc. 31, 1391 (2017); B. Goldsmith *et al.*, New J. Phys. 19, 013031 (2017)

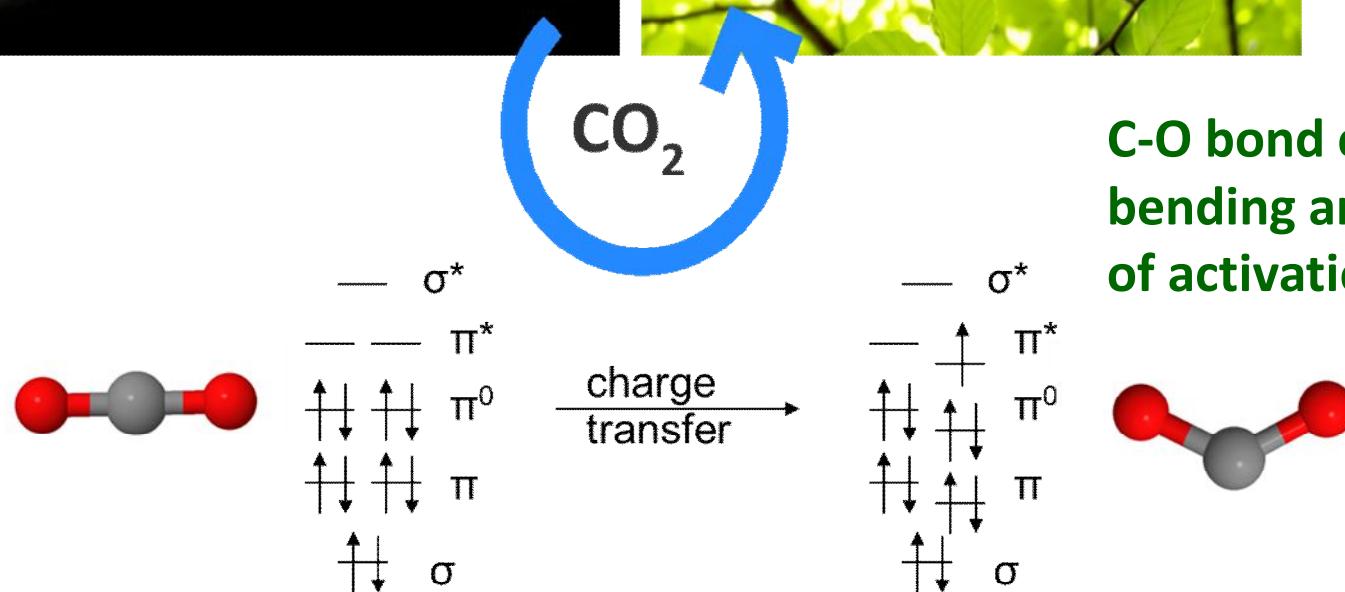
Data mining: Subgroup discovery



Data mining: Subgroup discovery

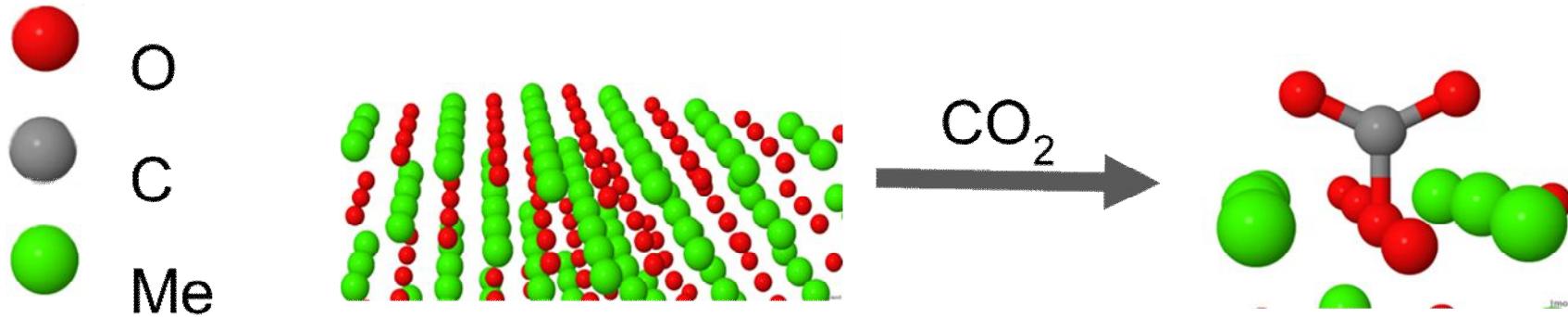


Subgroup discovery: CO₂ activation by adsorption



C-O bond elongation, O-C-O bending angle → indicators of activation

Subgroup discovery: CO₂ activation by adsorption



dry reforming of methane:
 $\text{CO}_2 + \text{CH}_4 = 2\text{H}_2 + 2\text{CO}$

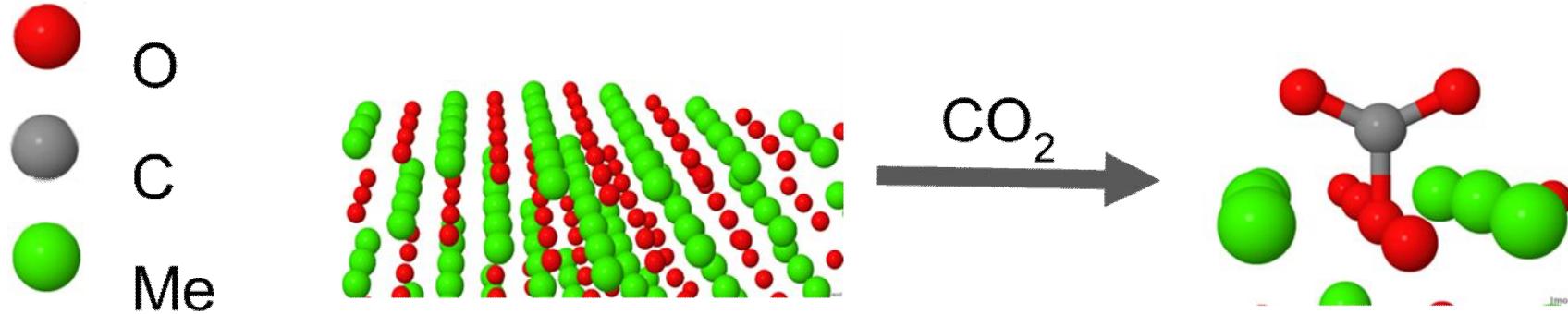
Sabatier reaction:
 $\text{CO}_2 + 4\text{H}_2 = \text{CH}_4 + 2\text{H}_2\text{O}$

partial hydrogenation:
 $\text{CO}_2 + 3\text{H}_2 = \text{CH}_3\text{OH} + \text{H}_2\text{O}$

Oxides:

- stable (structurally and compositionally) under increased temperatures;
- more resistant for poisoning;
- activation is frequently observed

Subgroup discovery: CO₂ activation by adsorption



C-O bond elongation, O-C-O bending angle → indicators of activation →

Which surface properties lead to desired indicators?

Use subgroup discovery to find materials that optimize activation indicators

$$f = N_{\text{subgroup}}/N_{\text{all}} \times (\text{mean}_{\text{subgroup}} - \text{mean}_{\text{all}}) \times (1 - \text{variance}_{\text{subgroup}}/\text{variance}_{\text{all}})$$

Maximize C-O bond length or O-C-O bending

Subgroup discovery: CO₂ activation by adsorption



| | | | | | | | | | | | | | | | | | |
|--------------------|--------------------|----------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| 1 H 1.008 | 2 | | | | | | | | | | | | | | 18 He 4.0026 | | |
| 3 Li 6.94 | 4 Be 9.0122 | | | | | | | | | | | | | | | | |
| 11 Na 22.990 | 12 Mg 24.305 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | |
| 19 K 39.098 | 20 Ca 40.078 | 21 Sc 44.956 | 22 Ti 47.867 | 23 V 50.942 | 24 Cr 51.996 | 25 Mn 54.938 | 26 Fe 55.845 | 27 Co 58.933 | 28 Ni 58.693 | 29 Cu 63.546 | 30 Zn 65.38 | 31 Ga 69.723 | 32 Ge 72.630 | 33 As 74.922 | 34 Se 78.97 | 35 Br 79.904 | 36 Kr 83.798 |
| 37 Rb 85.468 | 38 Sr 87.62 | 39 Y 88.906 | 40 Zr 91.224 | 41 Nb 92.906 | 42 Mo 95.95 | 43 Tc (98) | 44 Ru 101.07 | 45 Rh 102.91 | 46 Pd 106.42 | 47 Ag 107.87 | 48 Cd 112.41 | 49 In 114.82 | 50 Sn 118.71 | 51 Sb 121.76 | 52 Te 127.60 | 53 I 126.90 | 54 Xe 131.29 |
| 55 Cs 132.91 | 56 Ba 137.33 | 57-71 * | 72 Hf 178.49 | 73 Ta 180.95 | 74 W 183.84 | 75 Re 186.21 | 76 Os 190.23 | 77 Ir 192.22 | 78 Pt 195.08 | 79 Au 196.97 | 80 Hg 200.59 | 81 Tl 204.38 | 82 Pb 207.2 | 83 Bi 208.98 | 84 Po (209) | 85 At (210) | 86 Rn (222) |
| 87 Fr (223) | 88 Ra (226) | 89-103 # (265) | 104 Rf (268) | 105 Db (271) | 106 Sg (270) | 107 Bh (277) | 108 Hs (276) | 109 Mt (281) | 110 Ds (280) | 111 Rg (285) | 112 Cn (286) | 113 Nh (288) | 114 Fl (289) | 115 Mc (293) | 116 Lv (294) | 117 Ts (294) | 118 Og (294) |

* Lanthanide series

| | | | | | | | | | | | | | | |
|--------------------|--------------------|--------------------|--------------------|-------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| 57 La 138.91 | 58 Ce 140.12 | 59 Pr 140.91 | 60 Nd 144.24 | 61 Pm (145) | 62 Sm 150.36 | 63 Eu 151.96 | 64 Gd 157.25 | 65 Tb 158.93 | 66 Dy 162.50 | 67 Ho 164.93 | 68 Er 167.26 | 69 Tm 168.93 | 70 Yb 173.05 | 71 Lu 174.97 |
|--------------------|--------------------|--------------------|--------------------|-------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|

Actinide series

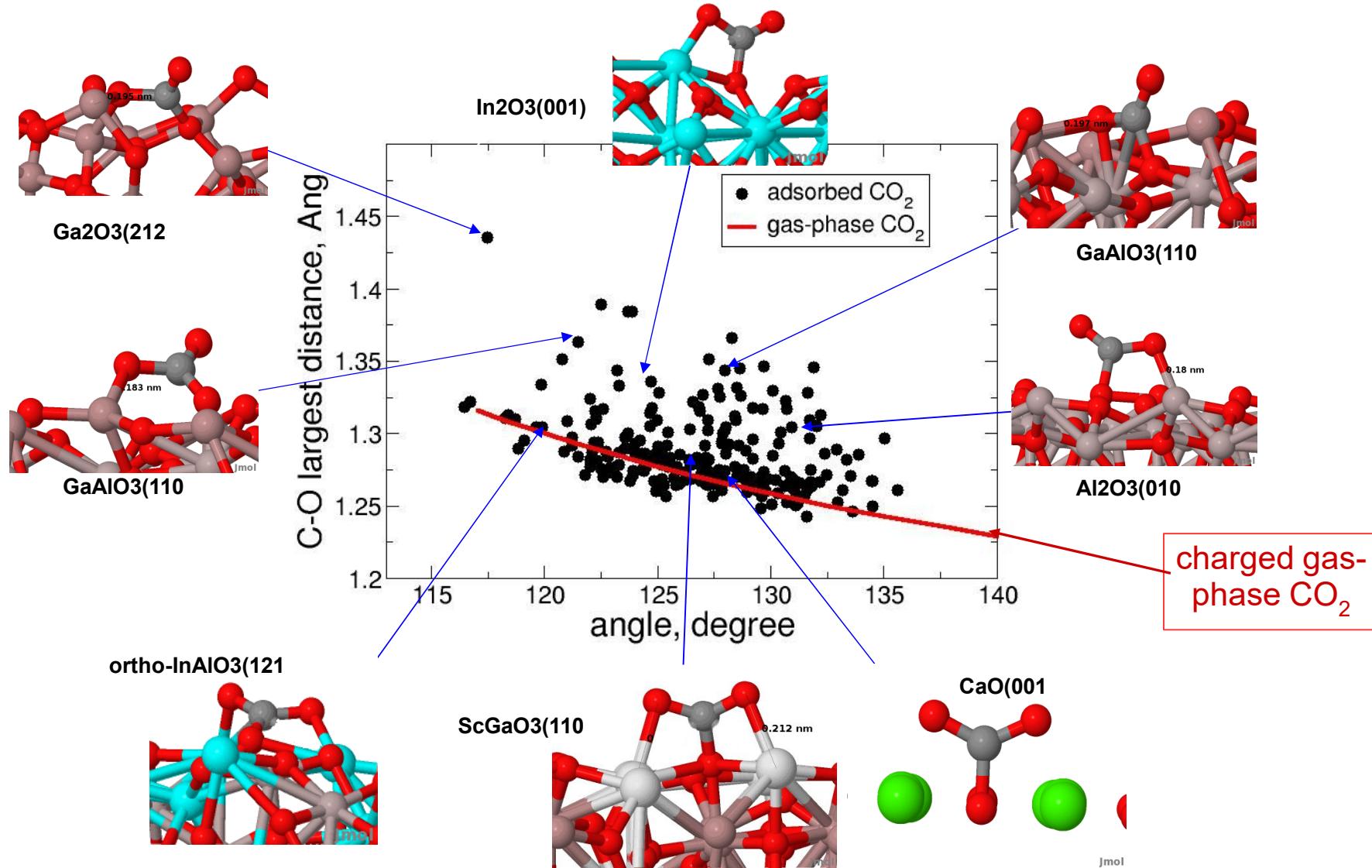
| | | | | | | | | | | | | | | |
|-------------------|--------------------|--------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|--------------------|--------------------|--------------------|--------------------|
| 89 Ac (227) | 90 Th 232.04 | 91 Pa 231.04 | 92 U 238.03 | 93 Np (237) | 94 Pu (244) | 95 Am (243) | 96 Cm (247) | 97 Bk (247) | 98 Cf (251) | 99 Es (252) | 100 Fm (257) | 101 Md (258) | 102 No (259) | 103 Lr (262) |
|-------------------|--------------------|--------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|--------------------|--------------------|--------------------|--------------------|

71 oxide materials

141 surfaces with Miller indexes ≤ 2

270 adsorption sites

Subgroup discovery: CO_2 activation by adsorption



Primary features

Atom:

electron affinity

$r_{l(\text{HOMO})}, r_{l-1}, r_{l+1}$

ionization potential

atomic numbers

electronegativity

Material:

work function

band gap

Cbm

surface form. energy

Site-specific features:

electrostatic potential

coordination number of O

distances to 1st, 2nd, 3^d nearest cations

Hirshfeld charge

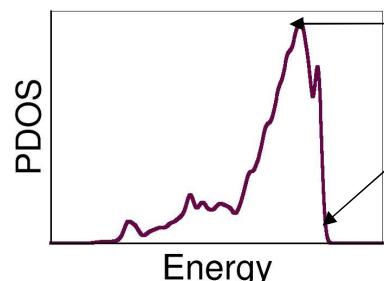
vdW C_6 -coefficient

local-structure parameters

bond-valence of O

polarizability

features of
O 2p-PDOS

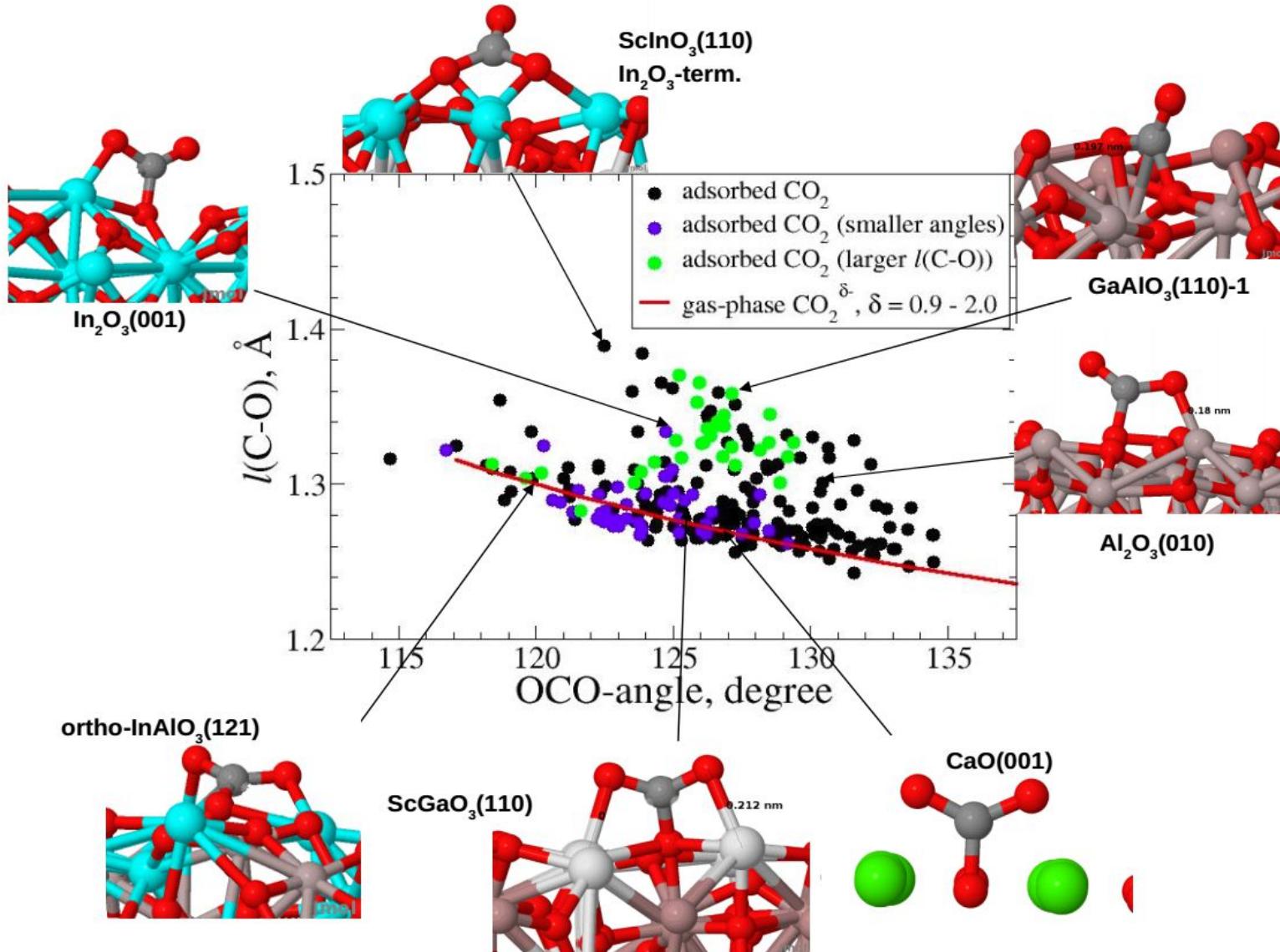


energy of maximum
energy of top

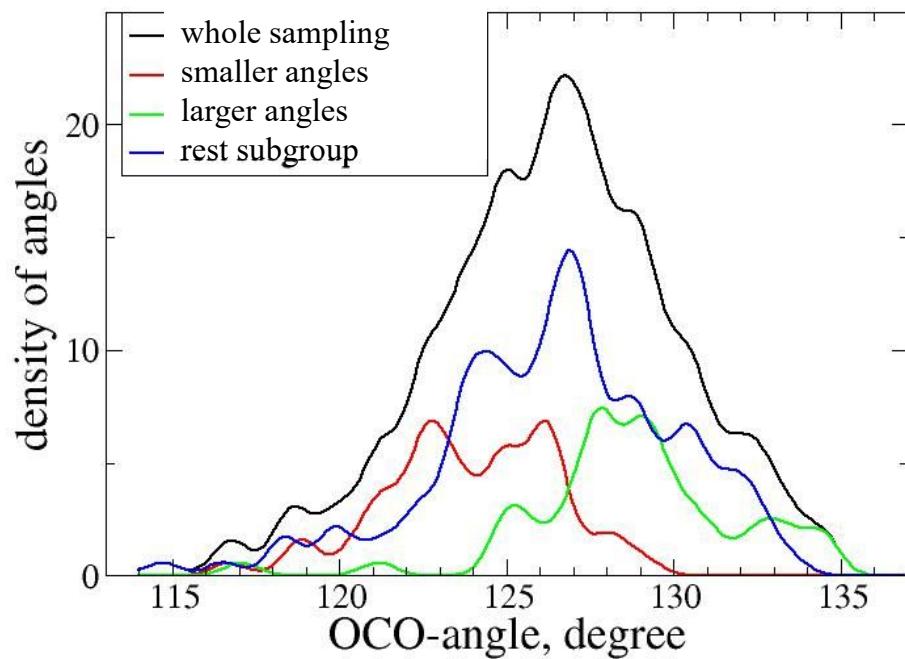
1st, 2nd, 3^d, 4th moments

DOS moments: center, width, skewness, kurtosis

Subgroup discovery: Adsorbed CO₂ properties

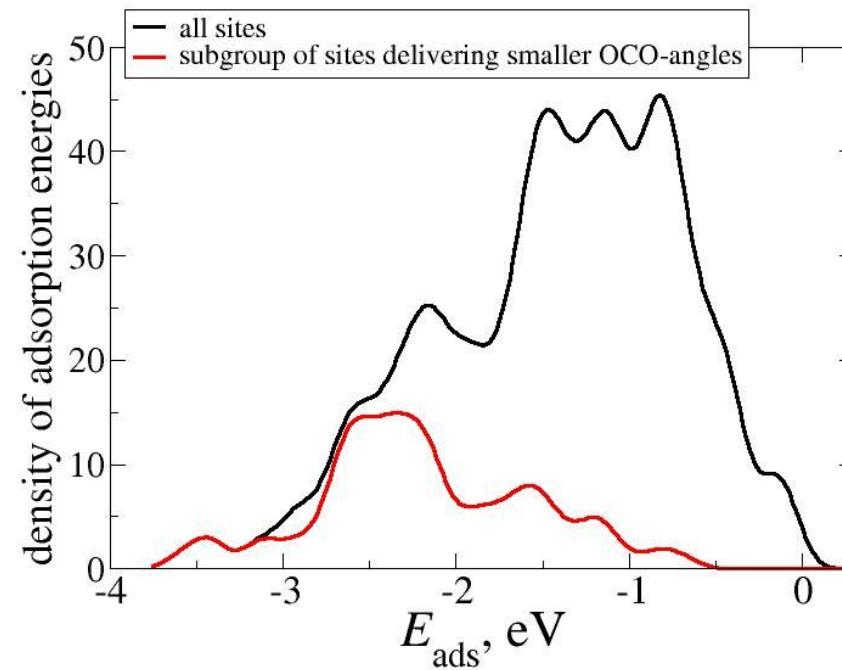


Subgroup discovery: Analysis of the OCO angle



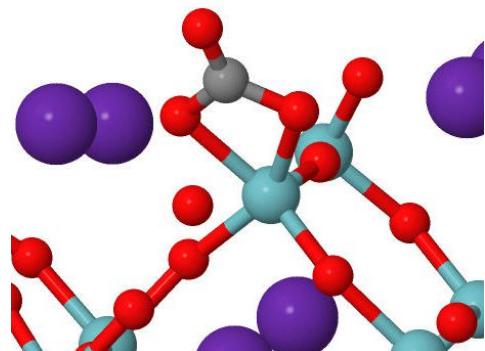
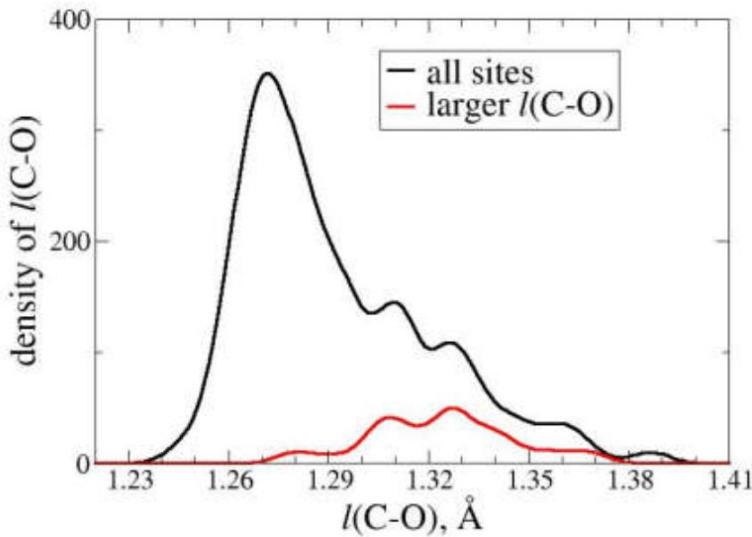
sites delivering smaller angles (59 adsorption sites):

(energy of O 2p band maximum > -6.0 eV) AND
(distance from O-site to first nearest cation > 1.8 Å) AND
(distance from O-site to second nearest cation > 2.1 Å)



Most of the site delivering smaller OCO angles are on ionic (basic) materials

Subgroup discovery: Analysis of the C-O bond length



sites delivering larger $l(\text{CO})$ (33 sites):

(cation charge < 0.5e) AND
(work function ≥ 5.2 eV) AND
(distance from O site to second nearest cation ≥ 2.14 Å)

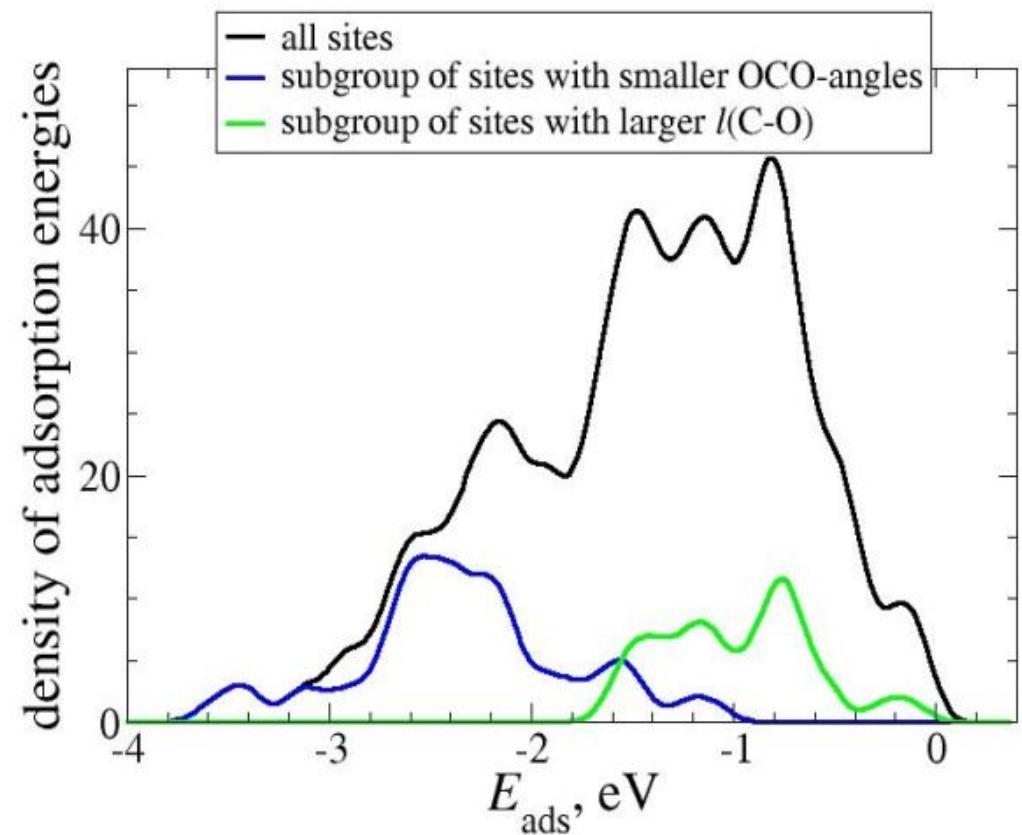
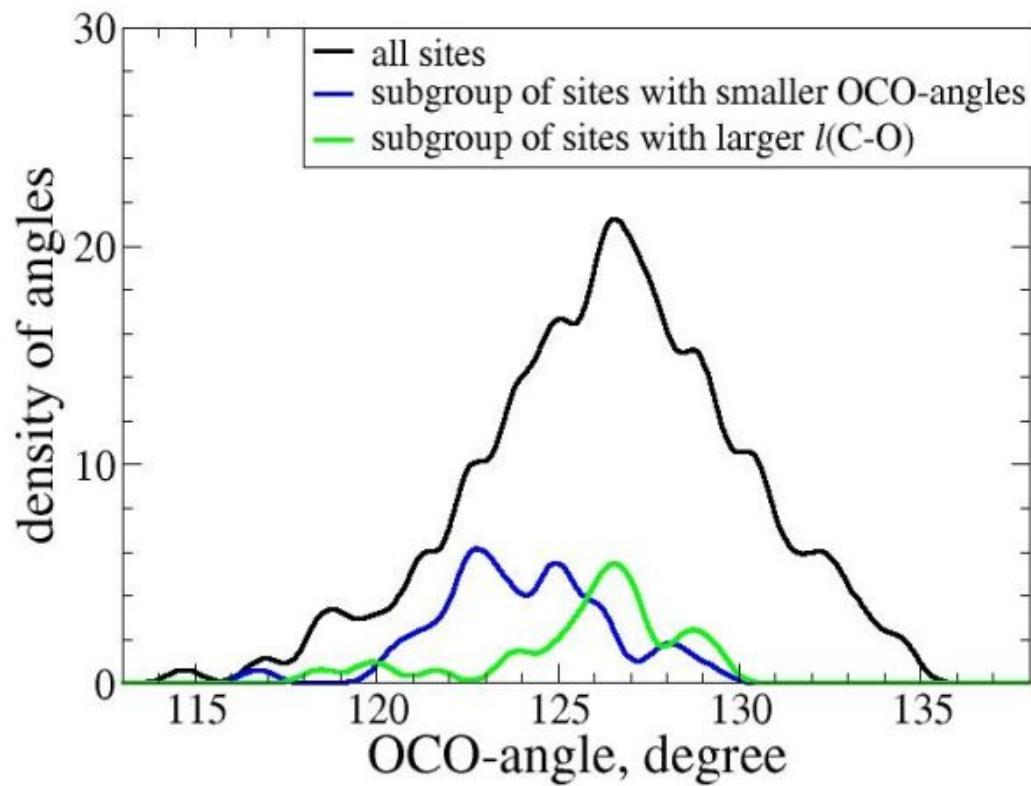
LaGaO_3 – cathode material in high-temperature electrochemical CO_2 reduction;

KNbO_3 – photocatalytic reduction of CO_2 into CH_4 ;

NaNbO_3 – photocatalyst for CO_2 reduction with $\sim 70\%$ of CO selectivity;

NaSbO_3 – material for CO_2 capture and storage (CCS)

Subgroup discovery: Alternative mechanisms of CO₂ activation



Longer C-O implies smaller OCO angles, but not too small → no catalyst poisoning

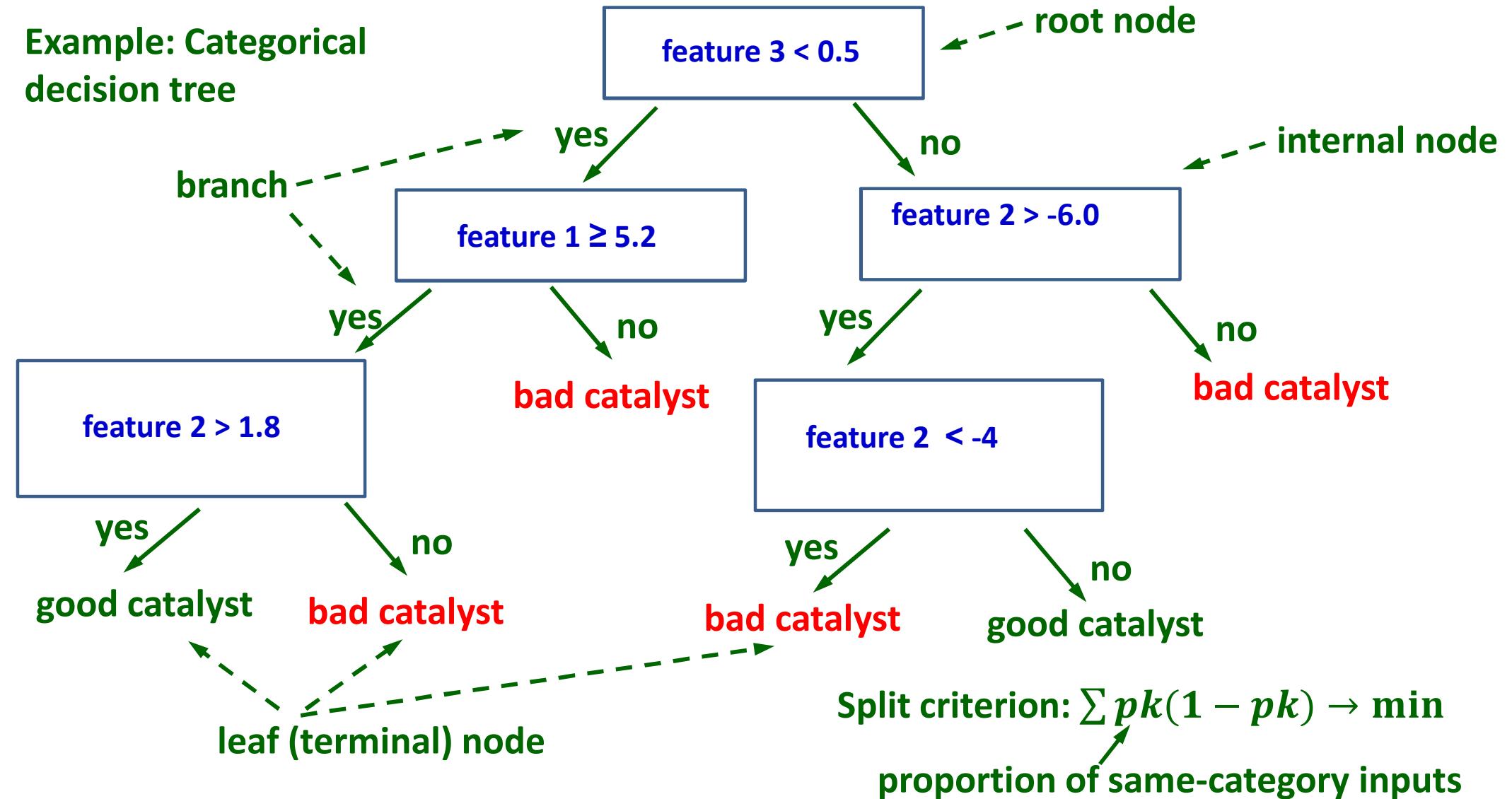
SISSO and SGD software

SISSO: <https://github.com/rouyang2017/SISSO>

Subgroup discovery: <https://bitbucket.org/realKD/creedo/wiki/Home>

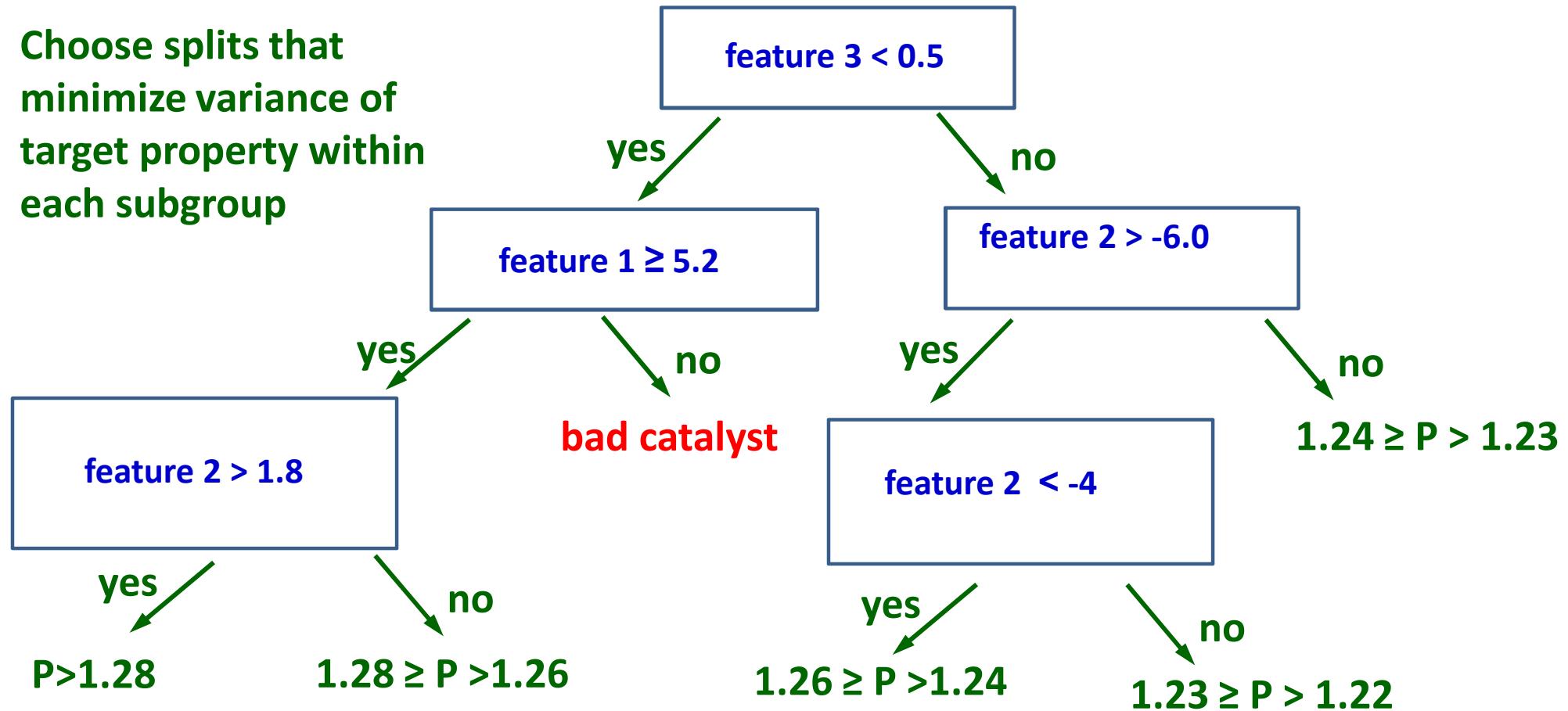
Decision trees

Example: Categorical decision tree



Decision tree regression

Choose splits that minimize variance of target property within each subgroup



Split criterion: $\sum(\text{target property} - \langle \text{target property} \rangle)^2 \rightarrow \min$ within each subgroup

Decision tree properties

- Simple to understand and interpret
- Global (important difference to subgroup discovery, which finds *locally unique* groups)
- Easy to overfit (can use LASSO-type penalty to solve this problem)
- Small change in data can lead to large change in the tree
- Relatively inaccurate

Random forest®

- 1) Perform tree regression or classification on several randomly selected subsets of data
- 2) In each tree, at each split choose randomly a fixed number of features, for which the best split is determined
- 3) Average predictions from the obtained trees

Properties:

- More accurate than a single tree (“each tree keeps other trees from making mistakes”)
- Interpretability of the model is lost
- Can be used to select primary features for other approaches such as SISSO

Random forest®

Interesting application: Identify most important surface structural features that determine surface stability



Automatic Prediction of Surface Phase Diagrams Using Ab Initio Grand Canonical Monte Carlo

Robert B. Wexler,[†] Tian Qiu,[†] and Andrew M. Rappe^{*,‡}



Chemical Pressure-Driven Enhancement of the Hydrogen Evolving Activity of Ni₂P from Nonmetal Surface Doping Interpreted via Machine Learning

Robert B. Wexler,[†] John Mark P. Martinez,[‡] and Andrew M. Rappe^{*,†}

Computational databases

General idea: Create infrastructure for storing, querying, and analyzing computational materials science data



The Materials Project

Harnessing the power of supercomputing and state of the art electronic structure methods, the Materials Project provides open web-based access to computed information on known and predicted materials as well as powerful analysis tools to inspire and design novel materials.

[Learn more](#)

 [Tutorials](#)

[Sign In or Register](#)

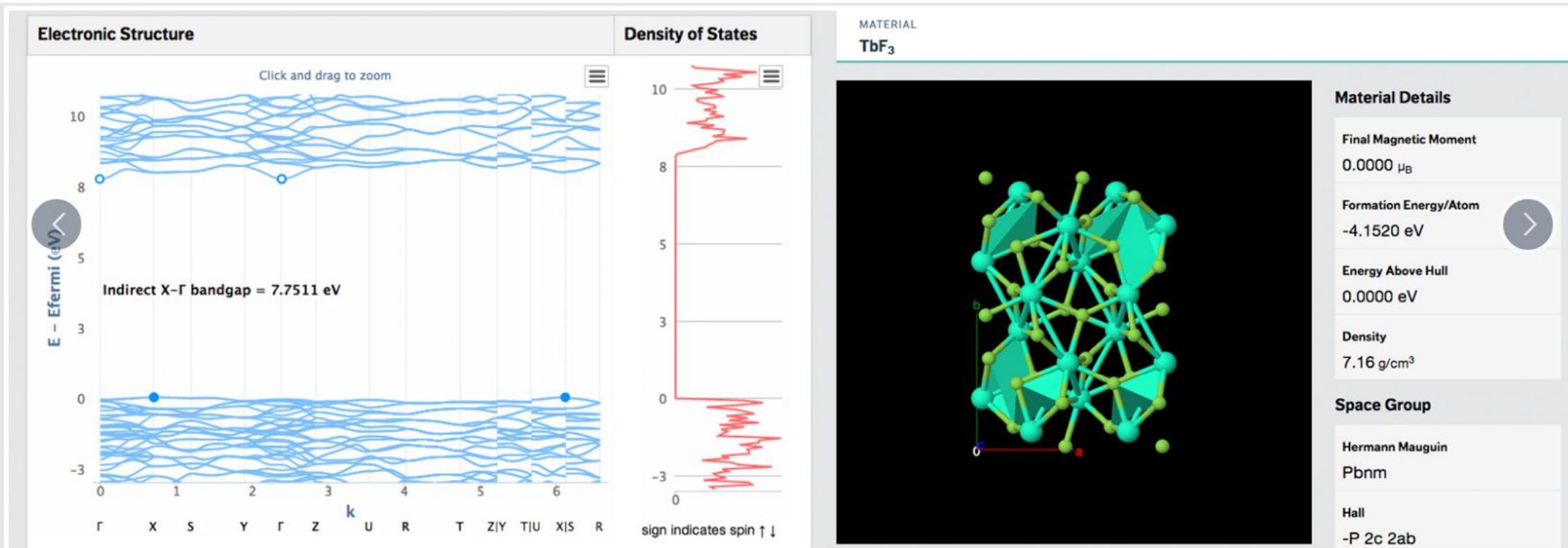
to start using

Leaders: Kristin Persson (Lawrence Berkeley National Laboratory), Gerbrand Ceder (University of California at Berkeley)

Structures are mostly from ICSD database (<https://icsd.products.fiz-karlsruhe.de/>)

Motto: We are in full control of the calculations and data. You can contribute, but first discuss with us

Materials Project: Features



EXPLORE MATERIALS

Search for materials information by chemistry, composition, or property

EXPLORE BATTERIES

Find candidate materials for lithium batteries. Get voltage profiles and oxygen evolution data.

VISUALIZE STABILITY

Generate phase and pourbaix diagrams to find stable phases and study reaction pathways

INVENT STRUCTURES

Design new compounds with our structure editor and substitution algorithms

CALCULATE

Calculate the enthalpy of 10,000+ reactions and compare with experimental values

Materials Project: Features

Database Statistics

131,613

INORGANIC COMPOUNDS

76,194

BANDSTRUCTURES

49,705

MOLECULES

530,243

NANOPOROUS MATERIALS

14,071

ELASTIC TENSORS

3,411

PIEZOELECTRIC TENSORS

4,730

INTERCALATION ELECTRODES

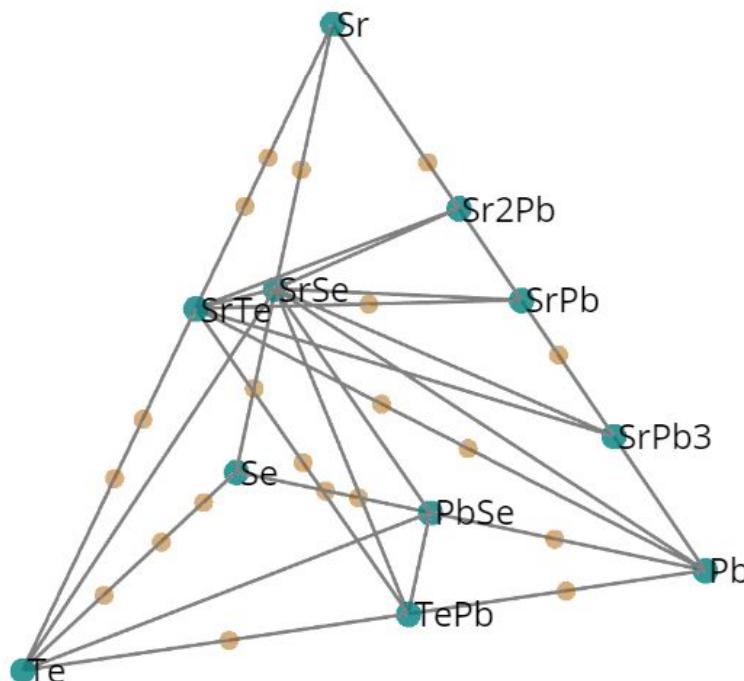
16,128

CONVERSION ELECTRODES

All calculations are performed with GGA or GGA+U

Typical data: relaxed crystal structure, band structure, DOS, energy from the convex hull, elastic properties, X-ray absorption and diffraction spectra, piezoelectric tensors,

The OQMD is a database of DFT calculated thermodynamic and structural properties of **815,654** materials, created in [Chris Wolverton's](#) group at Northwestern University.



Shortcuts

Search

Material Compositions

Query

Materials Data

Create

Phase Diagrams

Determine

Ground State Compositions (GCLP)

Visualize

Crystal Structures

RESTful API

OQMD API
Optimade API

The Open Quantum Materials Database: Features

Motto: We do all the calculations

All calculations are performed with GGA or GGA+U

Structures include hypothetical materials (not known experimentally)

Typical data: Formation and decomposition energies



AFLOW
Automatic - FLOW for Materials Discovery

HOME | CONSORTIUM | PUBLICATIONS | FORUM | SRC | SEARCH

AFLOW Schools and Seminars

AFLOW SCHOOL – Online

Welcome to AFLOW, a globally available database of **3,506,441** material compounds with over **694,275,318** calculated properties, and growing.

| | | | |
|-------------------------------|-------------------------------|-----------------------------|------------------------------|
| 348,185 band structures | 190,826 Bader charges | 6,174 elastic properties | 6,163 thermal properties |
| 1,715 binary systems | 362,989 binary entries | 30,203 ternary systems | 2,557,343 ternary entries |
| 150,623 quaternary systems | 450,635 quaternary entries | | |

AFLOW also offers online applications for property predictions using [machine learning](#), [prototype encyclopedia](#), and the generation of [convex hulls](#).

Automatic FLOW library: Features

Leader: Stefano Curtarolo (Duke University)

Motto: We do all the calculations

Calculations performed with GGA, GGA+U, ACBN0 (pseudo-hybrid)

Typical data: Relaxed geometries, electronic and **phonon band structures**, magnetic properties, thermodynamic properties

Provides tools for performing high-throughput calculations



Home

About

Pillars

Outreach

NOMAD Oasis

Events

NOMAD Centre of Excellence

Data Driven Science with an Overview Talk by Claudia Draxl

Feb 17, 2021 Symposium "Frontiers of Electroni

NOMAD'S KEY SERVICES CAN BE FOUND HERE:

The NOMAD (Novel Materials Discovery) Laboratory

A European Center of Excellence (CoE)

Leader: Matthias Scheffler (Humboldt University, Berlin)

**Motto: We will store your data, you decide if it is open access or restricted.
Even inaccurate calculations can be used for learning, provided all metadata (code version, method, basis set, etc.) are known**

Includes data from AFLOW, OQMD, Materials Project

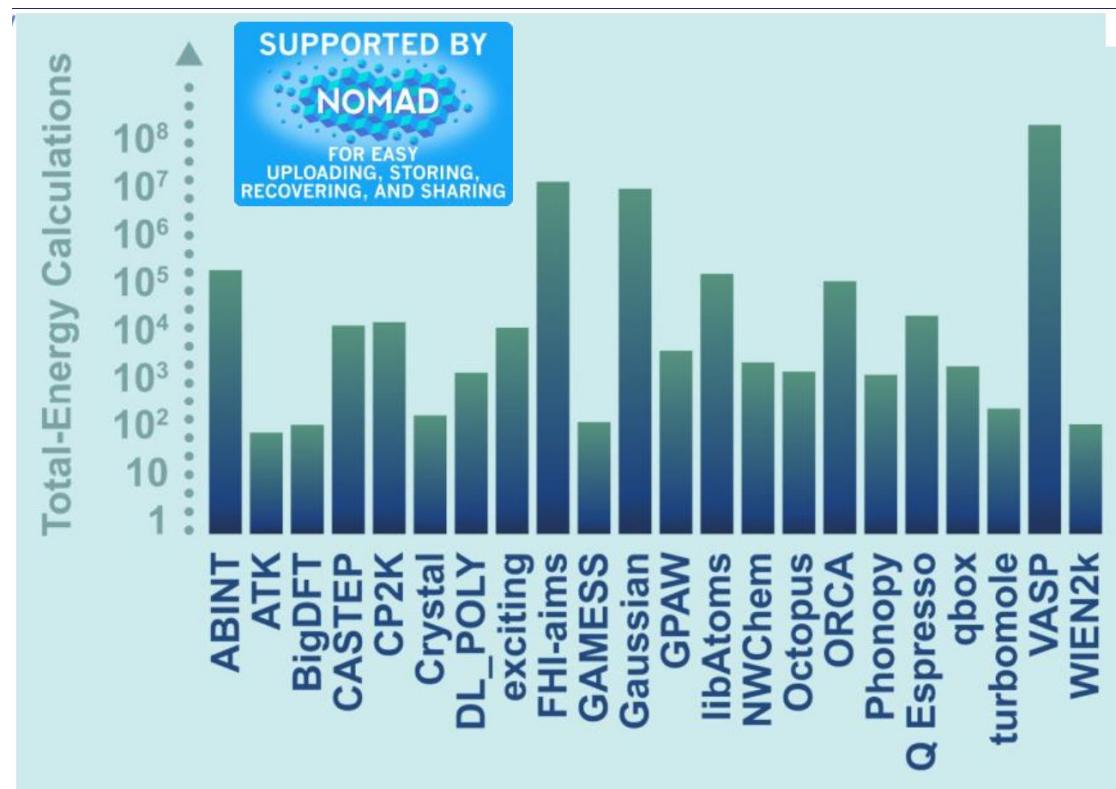
Automatic parsing of inputs and outputs from all major electronic-structure packages

Common format (metadata) for data from different electronic-structure packages

Parsable data: Total energies, geometry optimization, molecular dynamics, thermodynamic properties

The NOMAD (Novel Materials Discovery) Laboratory

A European Center of Excellence (CoE)

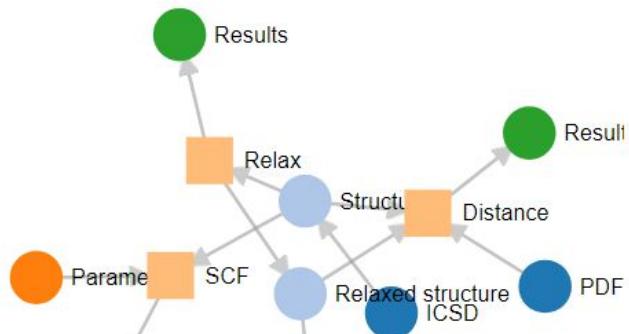


Contains raw data (input and output) uploaded by users



- [Workflows](#)
- [Data provenance](#)
- [Plugin framework](#)
- [HPC Interface](#)
- [Open Science](#)
- [Open source](#)

If you use AiiDA please cite:
AiiDA 1.0: S.P. Huber et al. arXiv:2003.12476 (2020)
AiiDA 0.x: G. Pizzi et al. Comp. Mat. Sci. 111, 218-230 (2016) (open access)



Most recent news

2020 Questionnaire results – AiiDA papers & testimonials

The results of the annual questionnaire on AiiDA-powered research projects are out! Find them on...

AiiDA v1.2.0 released

A new AiiDA release v1.2.0 is available! You can find more information at our download...

Pre-prints of upcoming AiiDA & Materials Cloud papers now available

After five years of continued development since the first AiiDA paper it was time to...

AiiDA at Google Summer of Code 2020

Thanks to the folks at NumFOCUS, AiiDA is participating in the Google Summer of Code...

AiiDA v1.1.1 released

A new AiiDA release v1.1.1 is available! You can find more information at our download...

Notes from AiiDA hackathon on plugin and workflow development

The AiiDA hackathon held at CINECA from February 17th-21st 2020 featured a number of presentations...

AiiDA v1.1.0 released

A new AiiDA release v1.1.0 is available! You can find more information at our download...

Automated Interactive Infrastructure and Database for Computational Science (AiiDA)

Leader: Nicola Marzari (EPFL, Switzerland)

Motto: Build your own database with our tools

Provides tools for performing high-throughput calculations