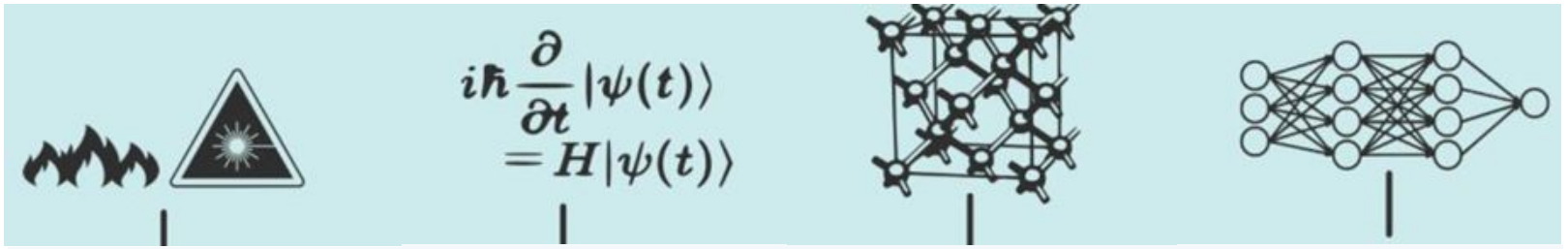
Data analytics approaches to materials design: Critical role of a descriptor

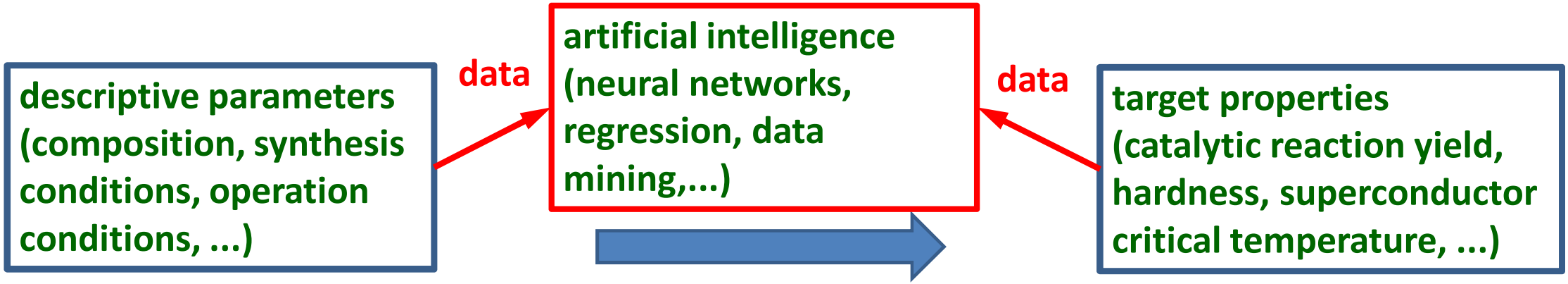
Sergey V. Levchenko

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

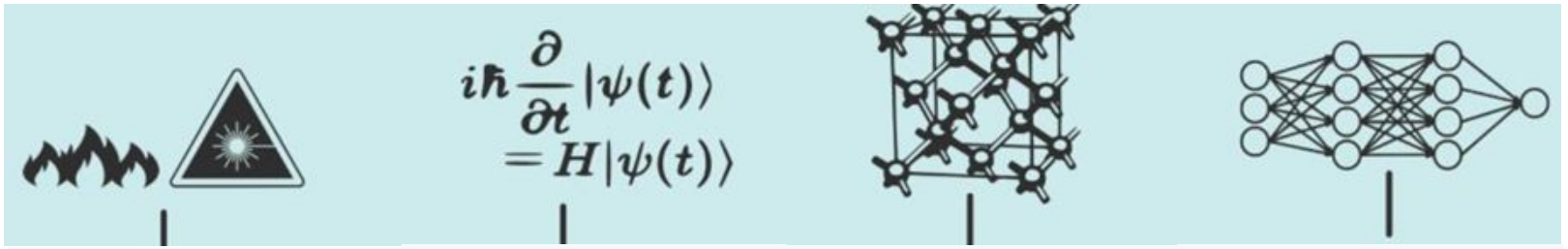
Research paradigm shift



Empirical science	Theoretical Science	Computational science	Data science
Experiments	Laws of classical mechanics, electrodynamics, etc.	Monte Carlo; molecular dynamics; density-functional theory and beyond	Detection of patterns and anomalies in Big Data; artificial intelligence; etc.



Research paradigm shift



Empirical science	Theoretical Science	Computational science	Data science
Experiments	Laws of classical mechanics, electrodynamics, etc.	Monte Carlo; molecular dynamics; density-functional theory and beyond	Detection of patterns and anomalies in Big Data; artificial intelligence; etc.

descriptive parameters (composition, synthesis conditions, operation conditions, ...)

data

artificial intelligence scikit-learn (https://scikit-learn.org/stable/)

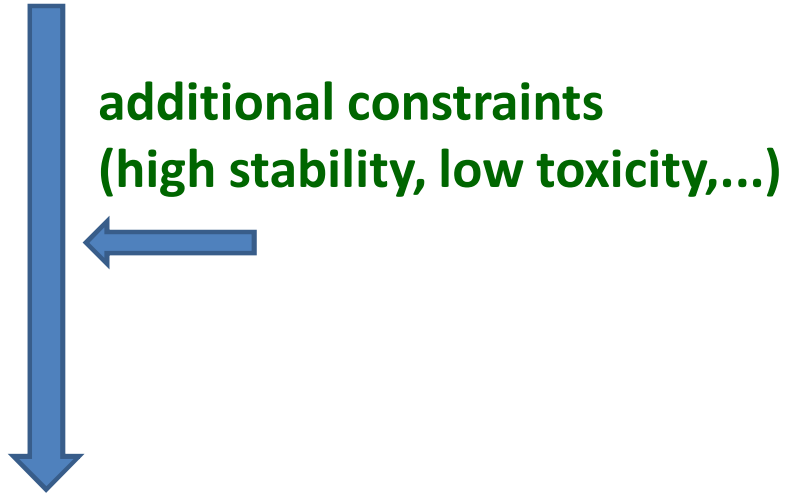
data

target properties (catalytic reaction yield, hardness, superconductor critical temperature, ...)

High-throughput computational materials design

Top-down design:

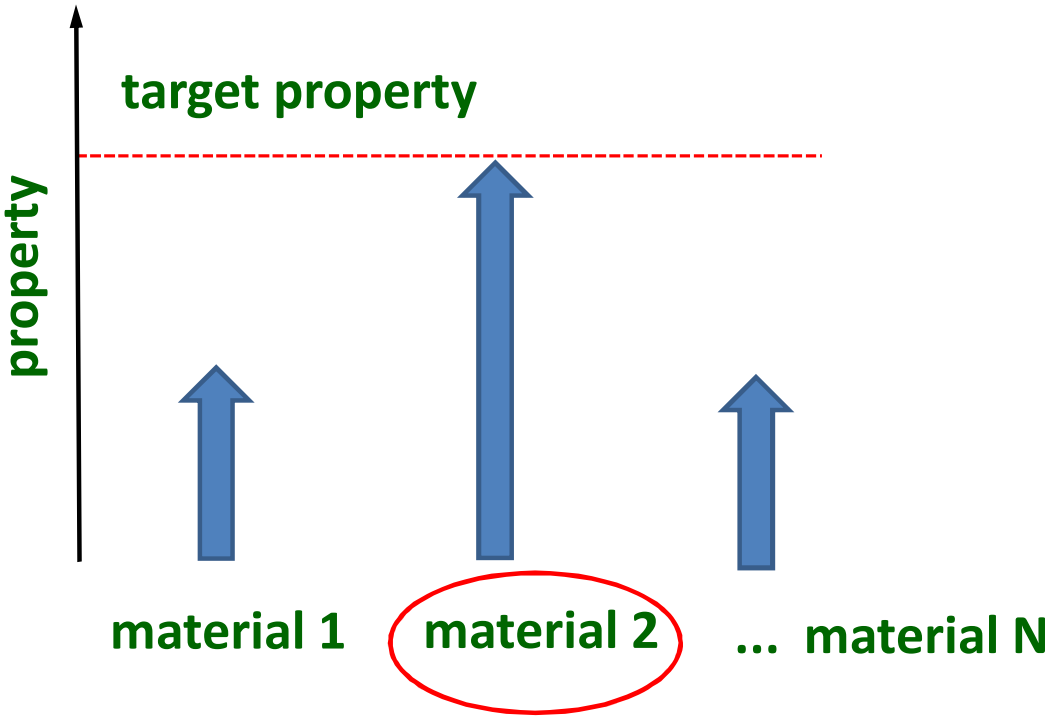
target property (high activity and selectivity of a catalyst)



synthesis recipe

not clear how to achieve this!

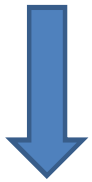
Bottom-up design:



high-throughput screening

The key issue: Complexity

$$i\frac{\partial\Psi(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n,\boldsymbol{R}_1,\boldsymbol{R}_2,\ldots,\boldsymbol{R}_N,t)}{\partial t}=\hat{H}(t)\Psi(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n,\boldsymbol{R}_1,\boldsymbol{R}_2,\ldots,\boldsymbol{R}_N,t)$$

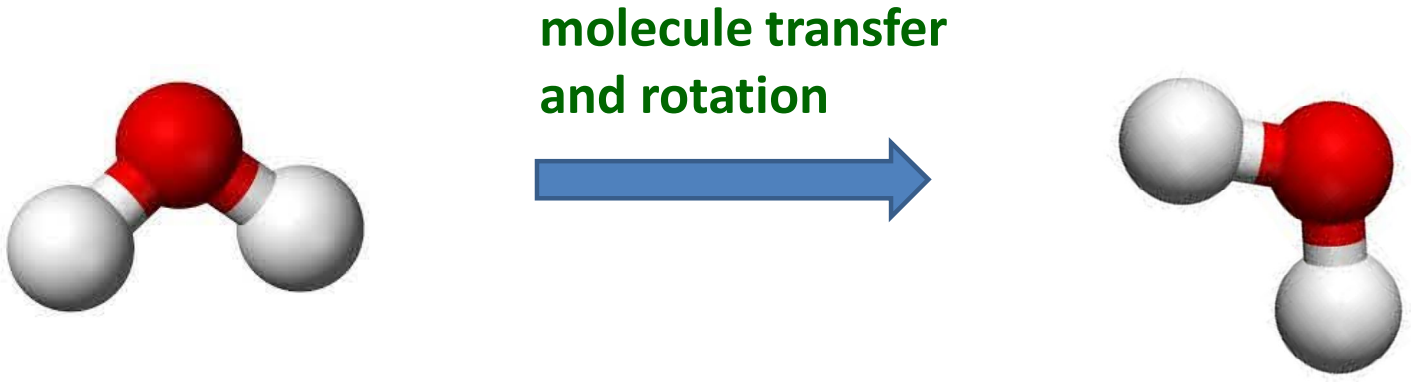


1) **Many-body problem ($3(n+N)$-dimensional)**

2) **Multiscale problem (tens orders of magnitude in time and space)**

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

Including science in descriptors



structure descriptor: Cartesian coordinates → changes, but properties do not change!

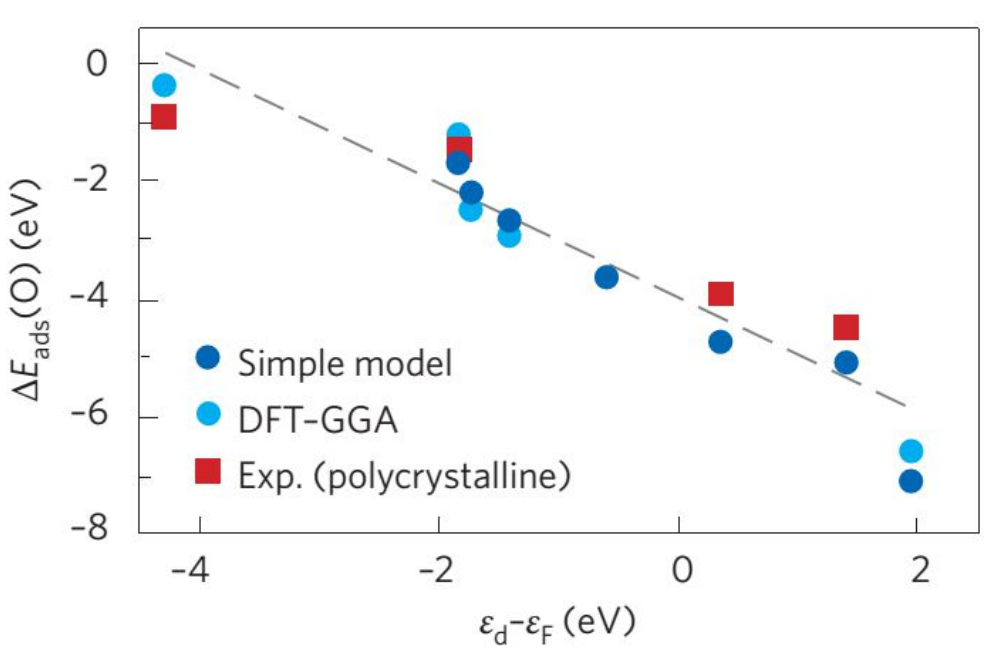
descriptive parameters (composition, synthesis conditions, operation conditions)

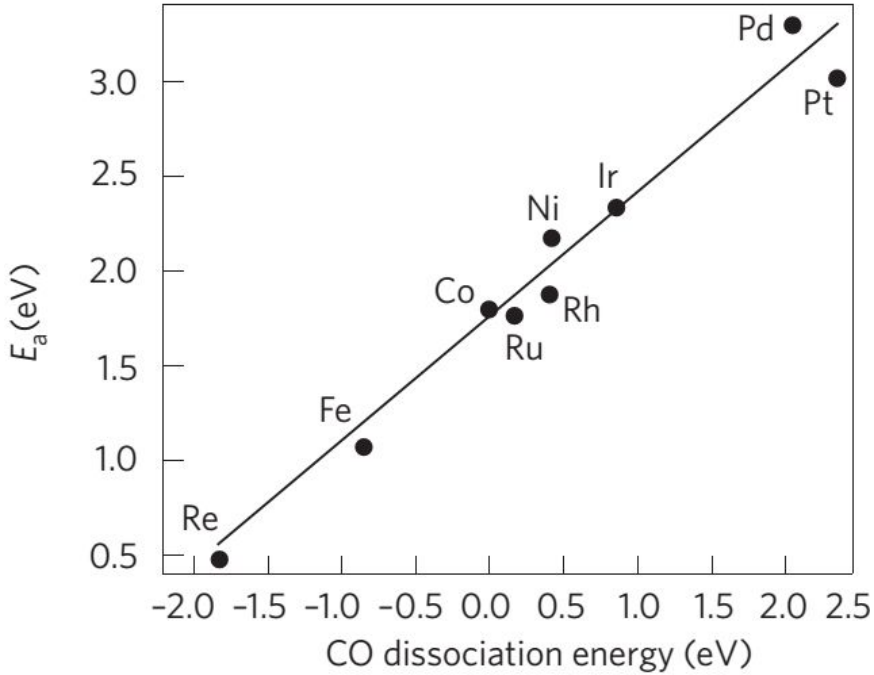
→

artificial intelligence (neural networks, regression, data mining,...)

machine will learn symmetries, not (other) physics -- much more data will be needed for an accurate model

Descriptors



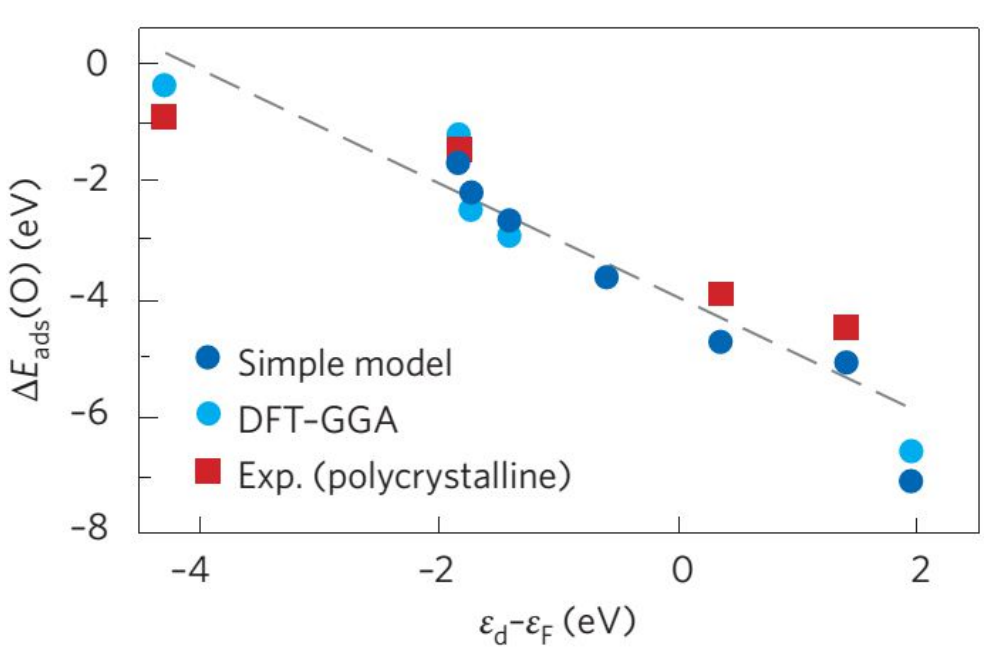


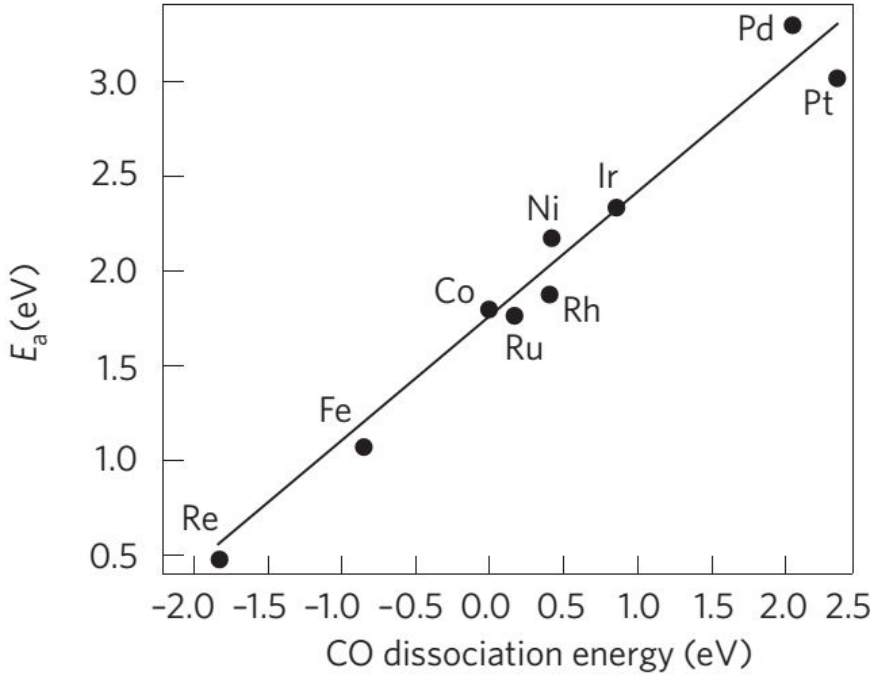
Simple(r) properties (bulk d-band center position and CO dissociation energy) are correlated to more complex properties (adsorption energy and reaction barrier)

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

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

Descriptors



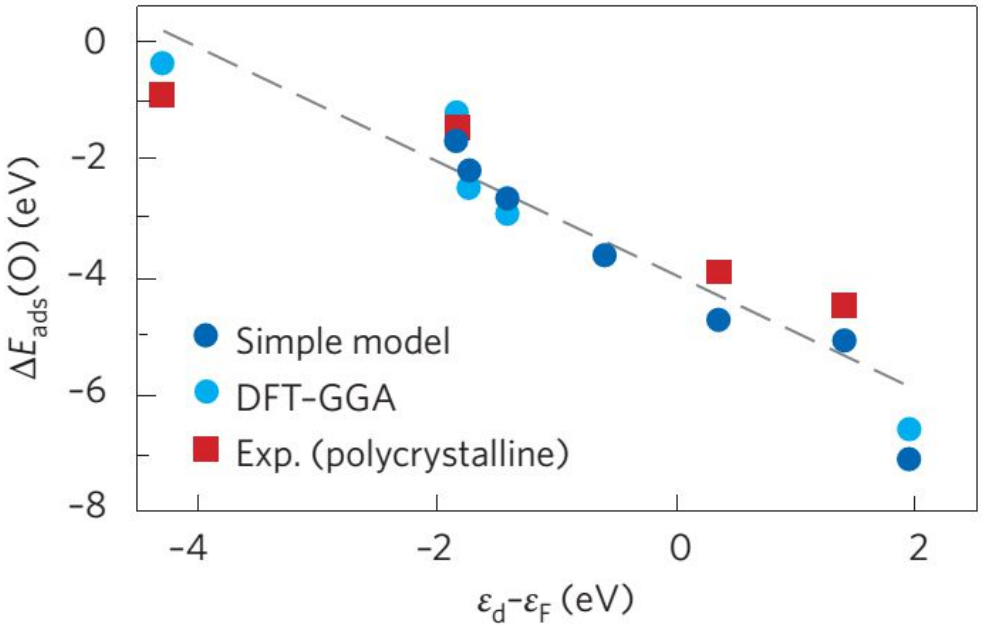


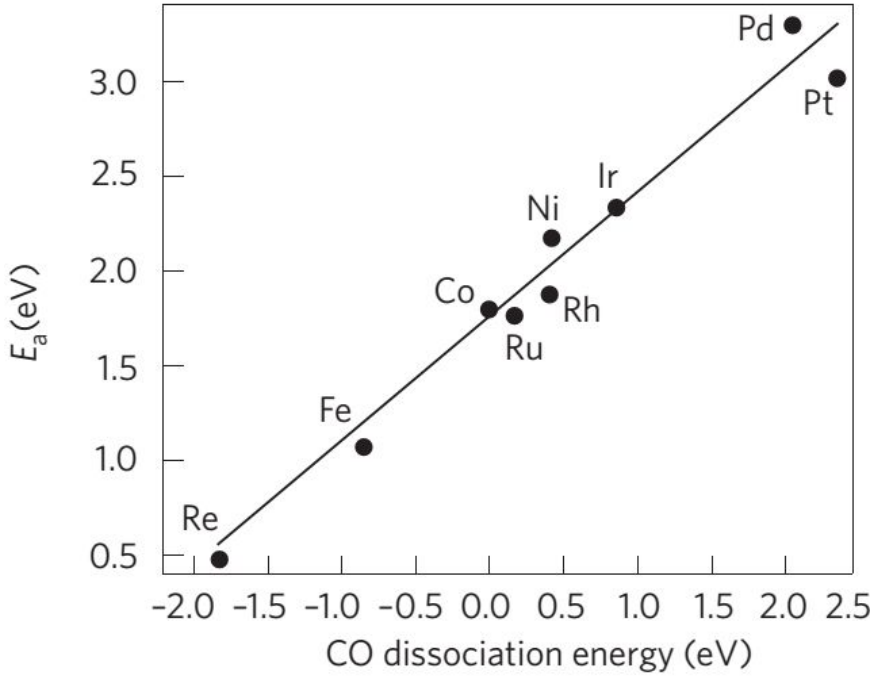
A simple physical model (Newns-Anderson) motivates the *d*-band center descriptor

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



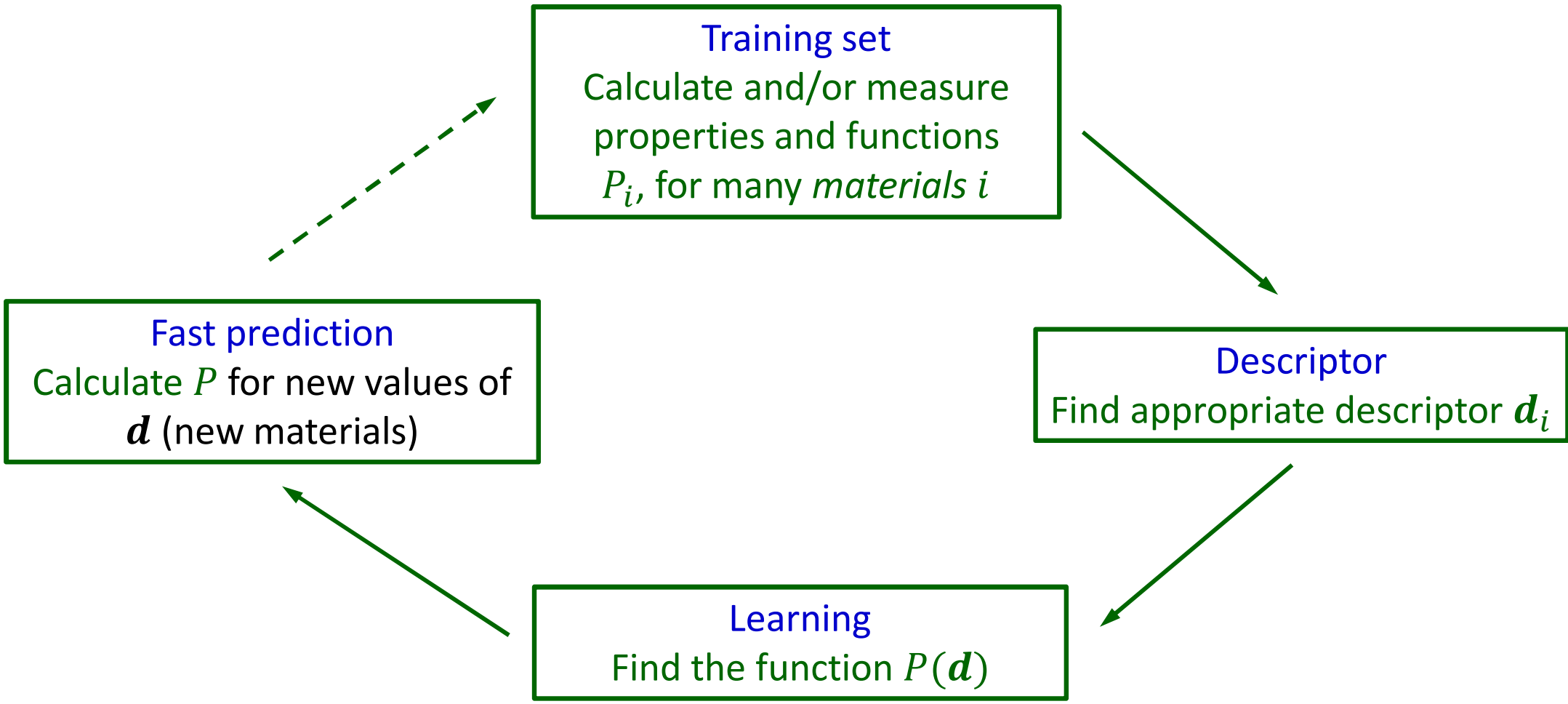


A simple physical model (Newns-Anderson) motivates the *d*-band center descriptor

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	linear
$P(\boldsymbol{d}) = \sum_{i=1}^{N} c_i \exp\left(-\|\boldsymbol{d}_i - \boldsymbol{d}\|_2^2 / 2\sigma^2\right)$	$P(\boldsymbol{d}) = \boldsymbol{d}\boldsymbol{c}$

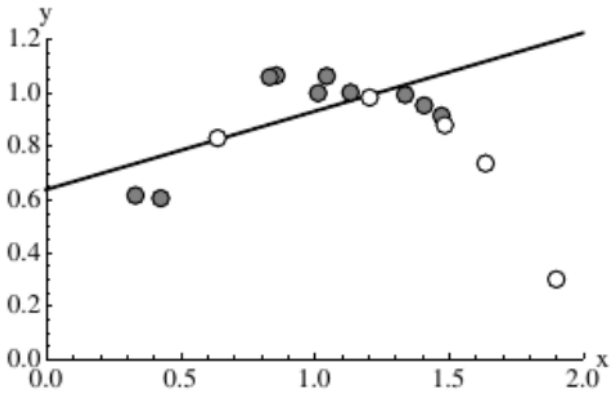
minimize

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

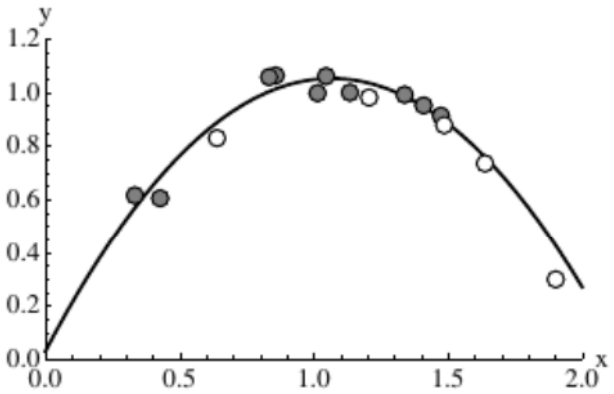
Regression: Importance of regularization

● training ○ validation

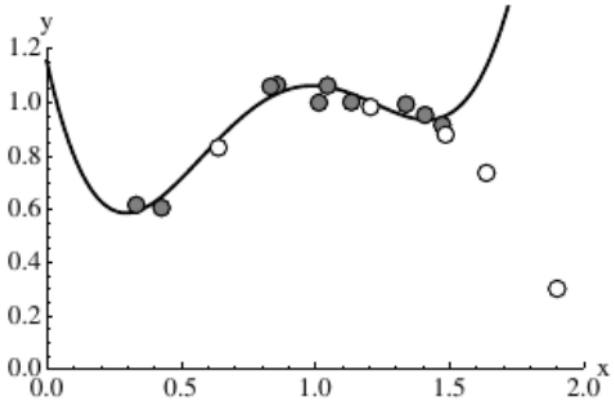
Underfitting



Fitting



Overfitting



Training/ validation error

0.123 / 0.443

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	linear
$P(\boldsymbol{d}) = \sum_{i=1}^{N} c_i \exp\left(-\|\boldsymbol{d}_i - \boldsymbol{d}\|_2^2 / 2\sigma^2\right)$	$P(\boldsymbol{d}) = \boldsymbol{d}\boldsymbol{c}$
minimize	
$\sum_{i=1}^{N} (P(\boldsymbol{d}_i) - P_i)^2 \quad +$	$\sum_{i=1}^{N} (P(\boldsymbol{d}_i) - P_i)^2 \; +$
$\lambda \sum_{i,j=1}^{N,N} c_i c_j \exp\left(-\|\boldsymbol{d}_i - \boldsymbol{d}_j\|_2^2 / 2\sigma^2\right)$	$\lambda \|\boldsymbol{c}\|_0$
$\|\boldsymbol{d}_i - \boldsymbol{d}_j\|_2^2 = \sum_{\alpha=1}^{\Omega} (d_{i,\alpha} - d_{j,\alpha})^2$	

Target property model: Kernel ridge regression versus feature selection

kernel (Gaussian, Laplacian, linear $(d_i \cdot d_j)$)

kernel ridge regression

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

$$\sum_{i=1}^{N} (P(\boldsymbol{d}_i) - P_i)^2 \quad +$$

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

penalty on similar data points

minimize

linear

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

$$\sum_{i=1}^{N} (P(\boldsymbol{d}_i) - P_i)^2 \quad +$$

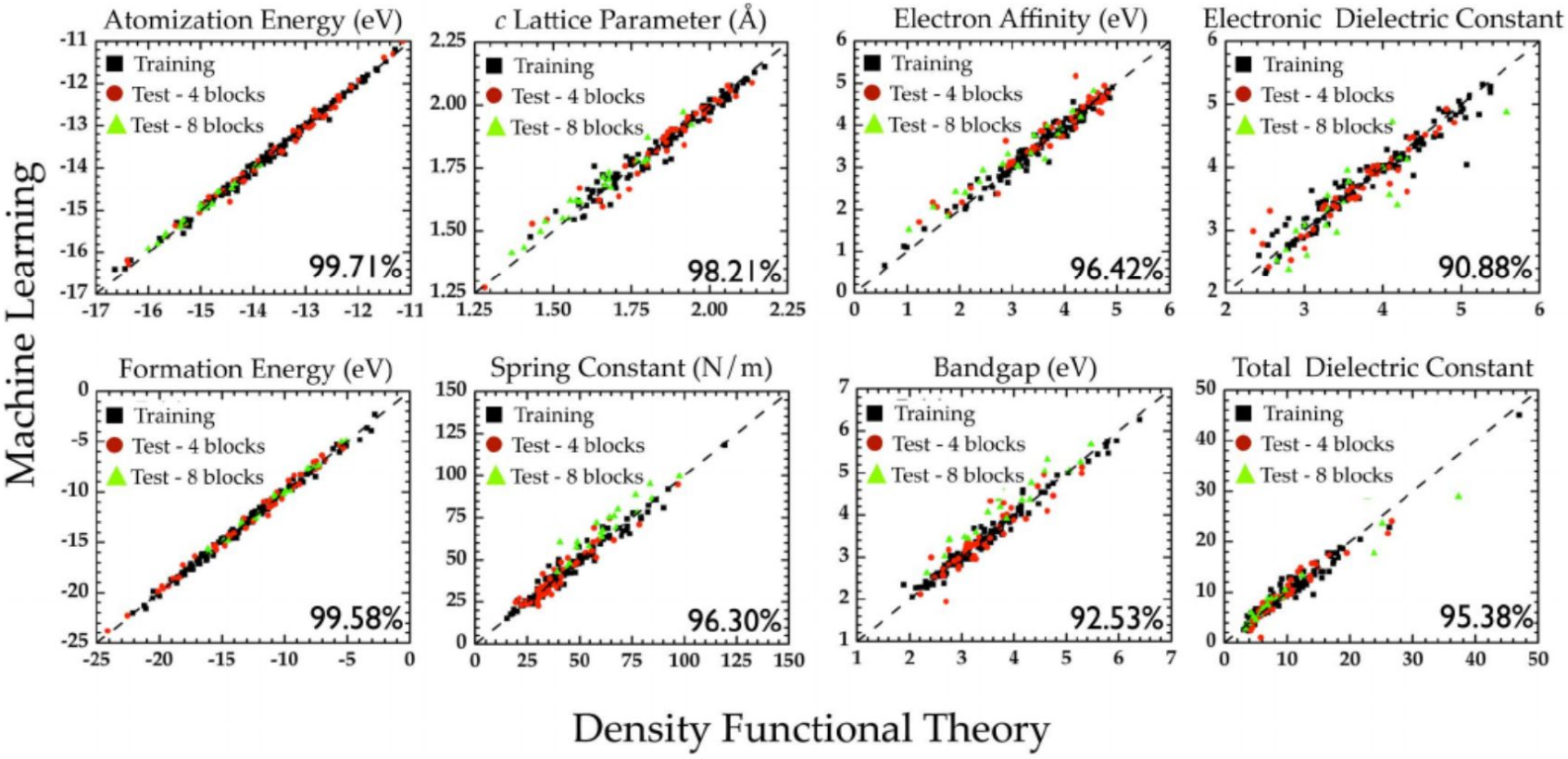
$$\lambda \|\boldsymbol{c}\|_0$$

penalty on the number of non-zero coefficients $\|\boldsymbol{c}\|_0$

(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]



Pilania, Wang, …, and Ramprasad, Scientific Reports 3, 2810 (2013). DOI: 10.1038/srep02810

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!

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

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

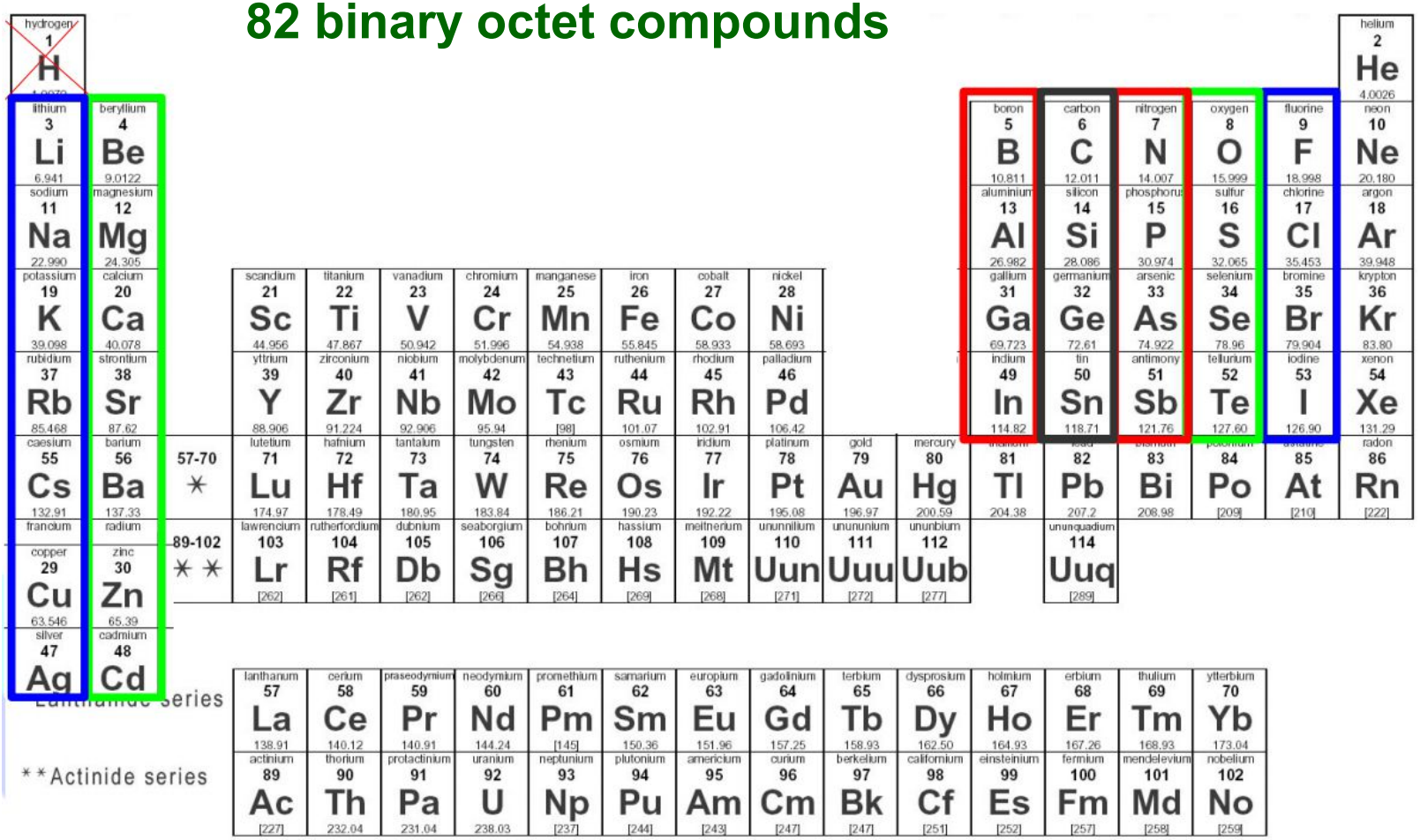
3) The dimension Ω of the descriptor should be as low as possible (for a certain accuracy request)

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

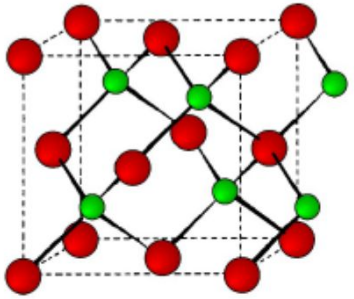
L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Phys. Rev. Lett. **114**, 105503 (2015)

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

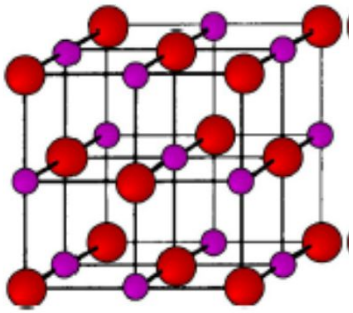
82 binary octet compounds



ZB



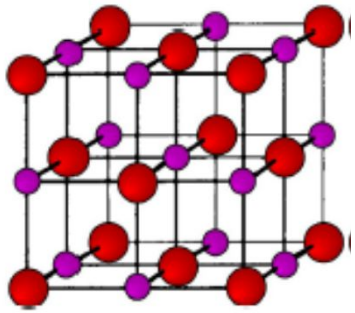
RS



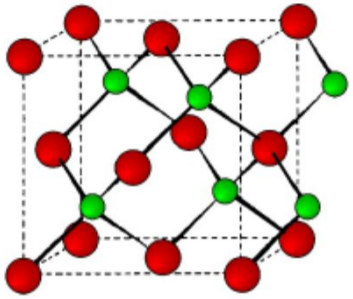
Proof of Concept: Descriptor for the Classification "Zincblende/Wurtzite (ZB/W) or Rocksalt (RS)?"

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

RS



ZB



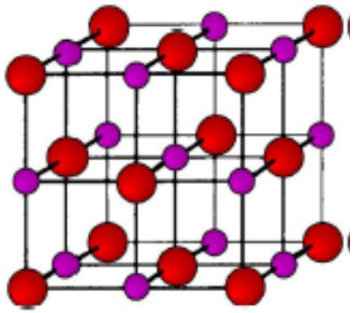
Energy differences between different structures are very small.

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

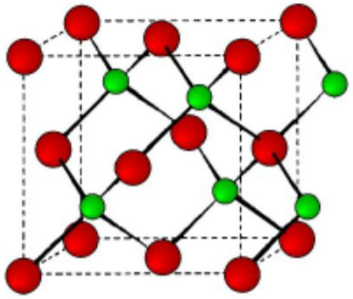
Proof of Concept: Descriptor for the Classification "Zincblende/Wurtzite (ZB/W) or Rocksalt (RS)?"

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

RS



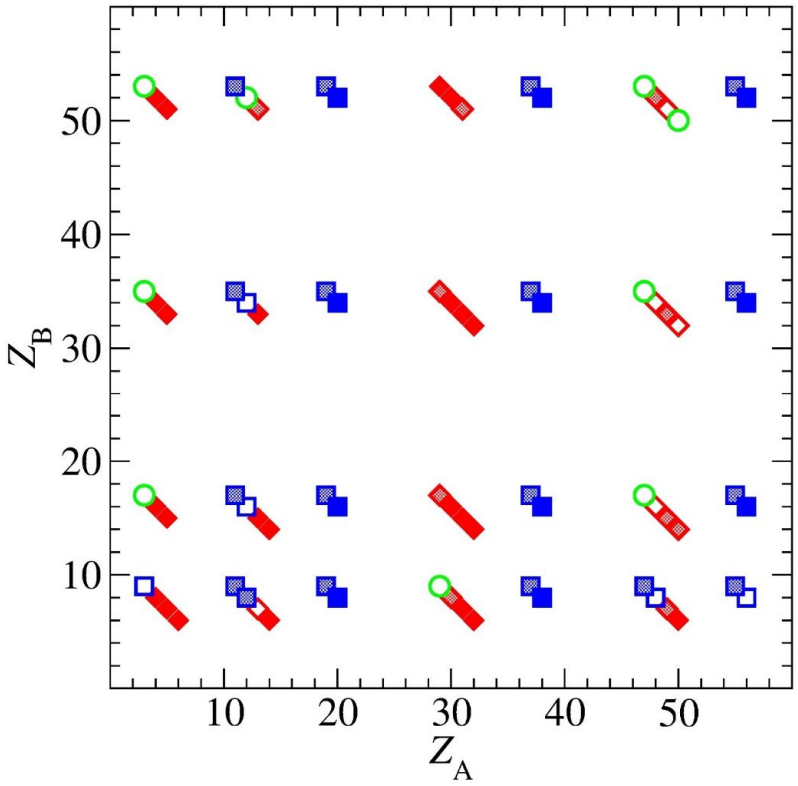
ZB

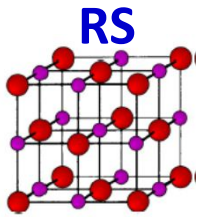


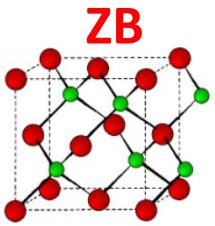
J. A. van Vechten, Phys. Rev. 182, 891 (1969). **J. C. Phillips**, Rev. Mod. Phys. 42, 317 (1970).
J. John and A.N. Bloch, Phys. Rev. Let. 33, 1095 (1974) **J. R. Chelikowsky and J. C. Phillips**, Phys. Rev. B 33, 2453 (1978)
A. Zunger, Phys. Rev. B 22, 5839 (1980).
D. G. Petifor, Solid State Commun. 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?"**







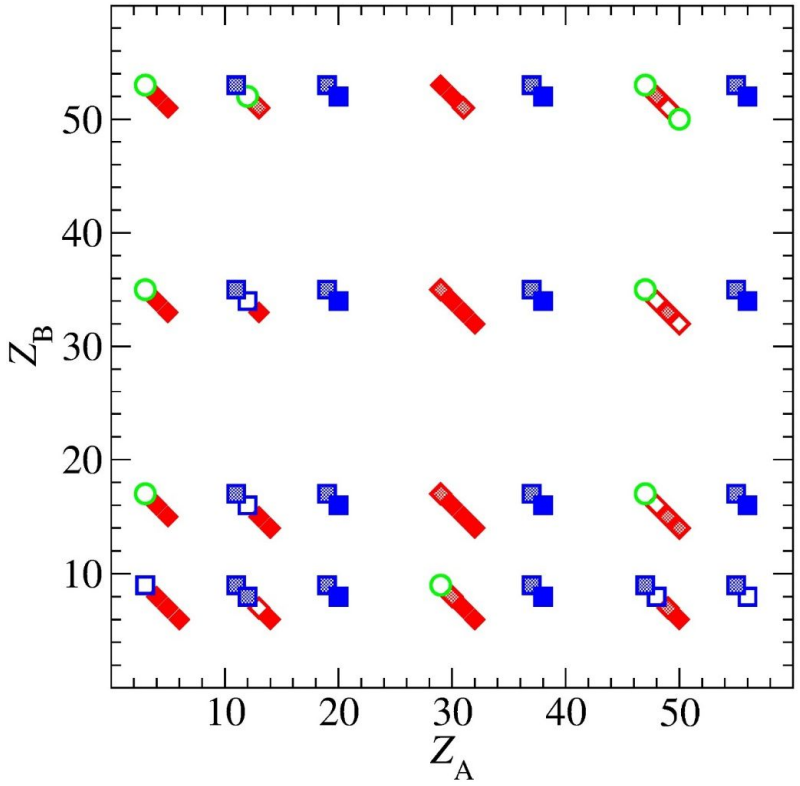
$\Delta = E(\mathrm{RS}) - E(\mathrm{ZB})$

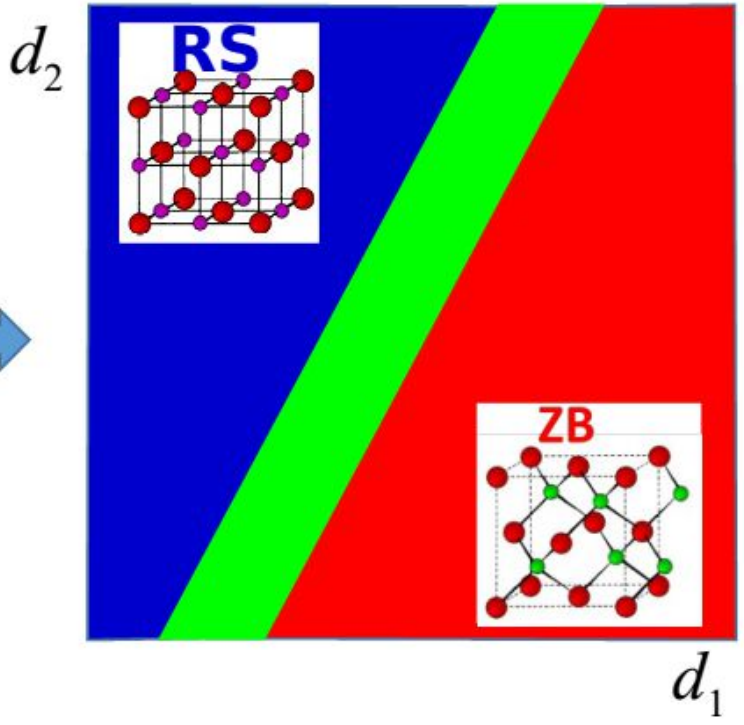
- ZB, $\Delta > 0.2$ eV
- ZB, $0.1\ \mathrm{eV} < \Delta \leq 0.2\ \mathrm{eV}$
- ZB, $0.05\ \mathrm{eV} < \Delta \leq 0.1\ \mathrm{eV}$
- $-0.05\ \mathrm{eV} < \Delta \leq 0.05\ \mathrm{eV}$
- RS, $-0.1\ \mathrm{eV} < \Delta \leq -0.05\ \mathrm{eV}$
- RS, $-0.2\ \mathrm{eV} < \Delta \leq -0.1\ \mathrm{eV}$
- RS, $\Delta \leq -0.2\ \mathrm{eV}$

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 lives here and the *RS* community there?"**

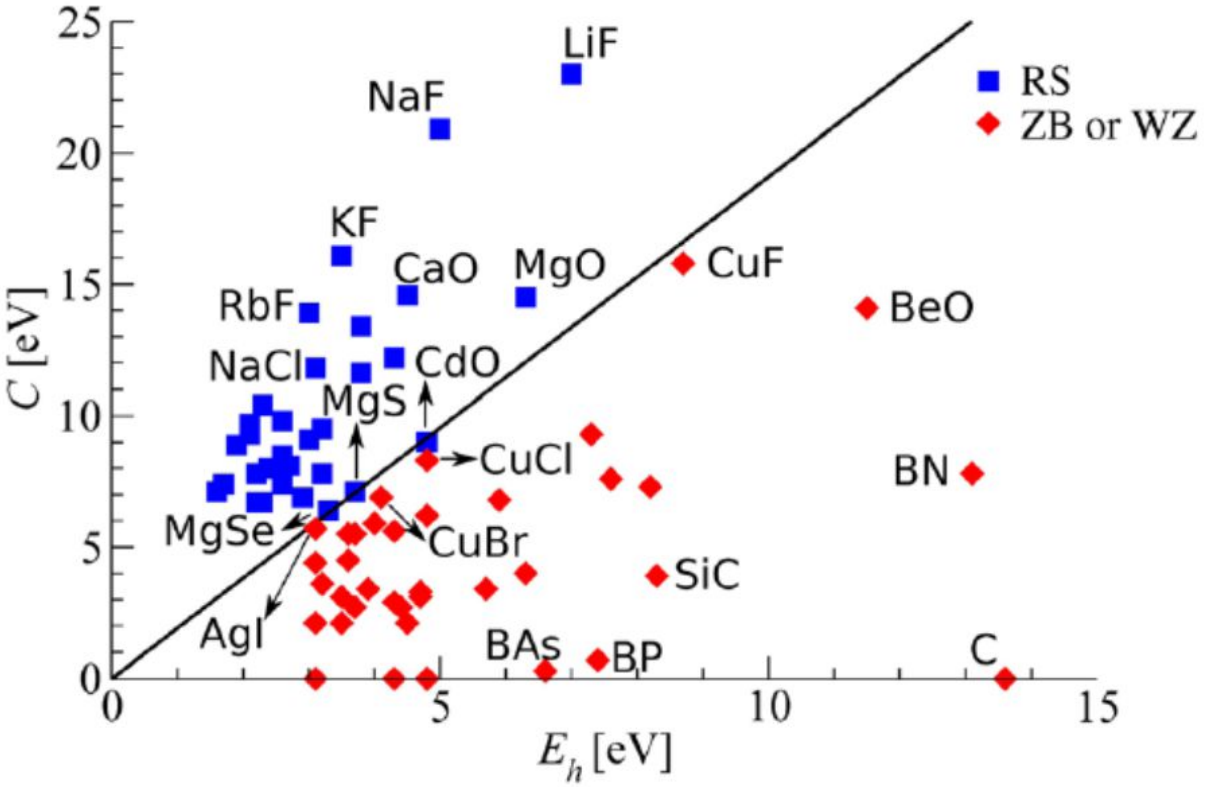
No complexity reduction → need a better basis





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?"**



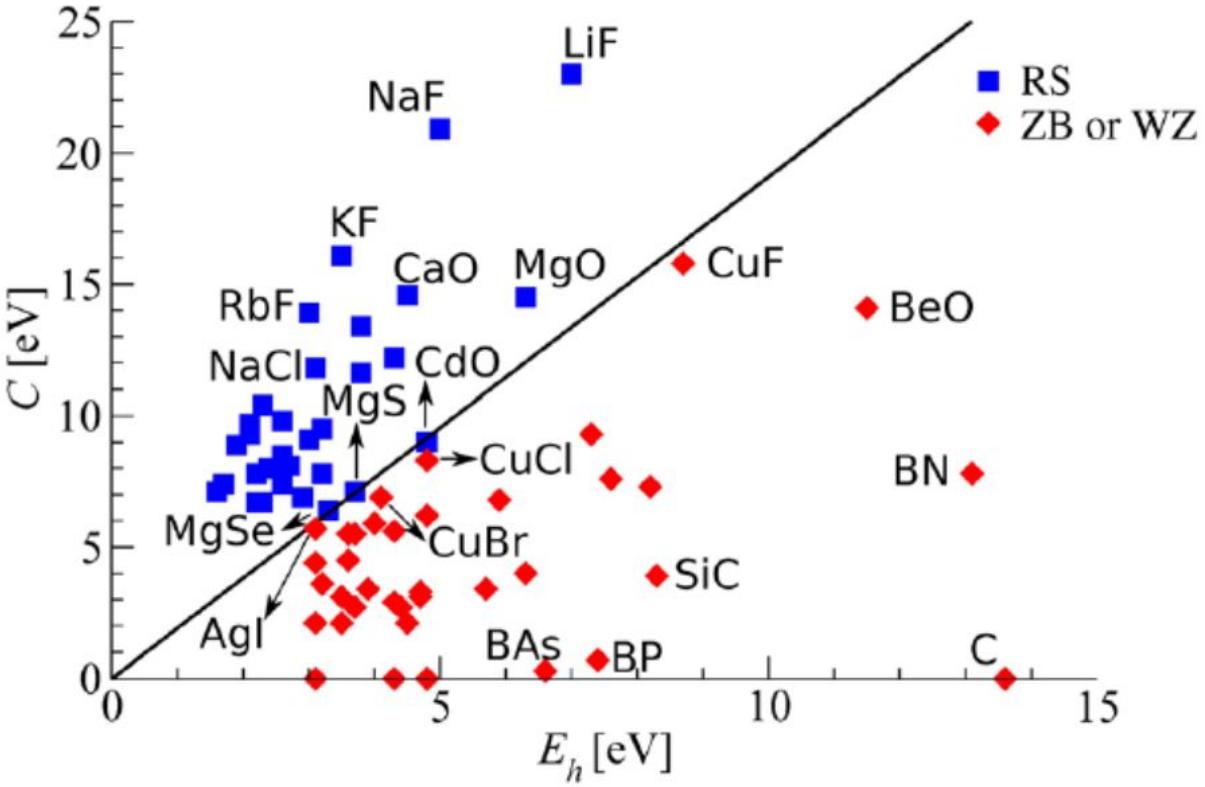
descriptor can be determined spectroscopically - properties of the solid

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. Let. 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

free atoms

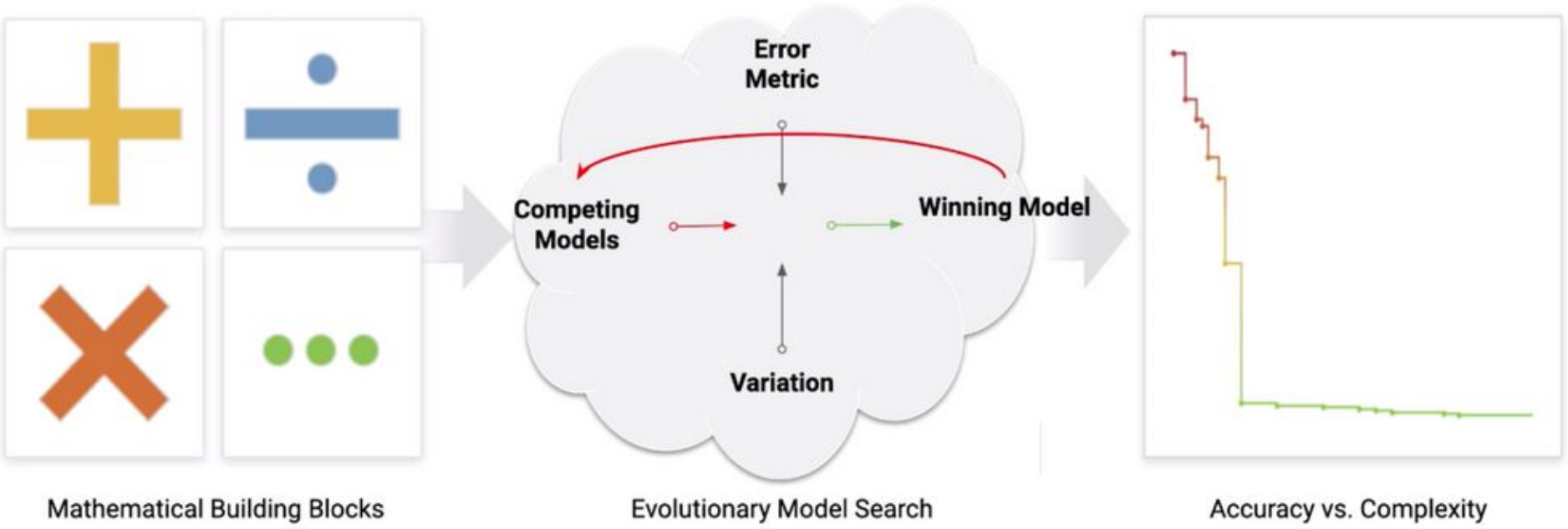
ID	Description	Symbols	#
$A1$	Ionization Potential (IP) and Electron Affinity (EA)	IP(A) EA(A) IP(B) EA(B) [1]	4
$A2$	Highest occupied (H) and lowest unoccupied (L) Kohn-Sham levels	H(A) L(A) H(B) L(B)	4
$A3$	Radius at the max. value of s, p, and d valence radial radial probability density	r_s(A) r_p(A)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

free dimers

ID	Description	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

https://community.datarobot.com/t5/resources/introduction-to-eureqa/ta-p/2409

Primary features and feature space

free atoms

ID	Description	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 s, p, and d valence radial radial probability density	r_s(A) r_p(A) r_d(A) r_s(B) r_p(B) r_d(B)	6

free dimers

ID	Description	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 $A1$	$\lvert\text{IP(A)} \pm \text{IP(B)}\rvert$	12
$B2$	absolute differences and sums of $A2$	$\lvert\text{L(B)} \pm \text{H(A)}\rvert$	12
$B3$	absolute differences and sums of $A3$	$\lvert r_p(\text{A}) \pm r_s(\text{A})\rvert$	30
$C3$	squares of $A3$ and $B3$ (only sums)	$r_s(\text{A})^2, (r_p(\text{A}) + r_s(\text{A}))^2$	21
$D3$	exponentials of $A3$ and $B3$ (only sums)	$\exp(r_s(\text{A})), \exp(r_p(\text{A}) \pm r_s(\text{A}))$	21
$E3$	exponentials of squared $A3$ and $B3$ (only sums)	$\exp(r_s(\text{A})^2), \exp(r_p(\text{A}) \pm r_s(\text{A})^2)$	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., $\left|r_s(A_j) - r_p(B_j)\right|$) (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?

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

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., $\left|r_s(A_j) - r_p(B_j)\right|$) (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?

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

$\|c\|_0$ -- number of non-zero coefficients → 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:

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

How to find c_l?

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

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

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

$\|c\|_1 = \sum_l |c_l|$ -- LASSO (Least Absolute Shrinkage and Selection Operator) → 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)

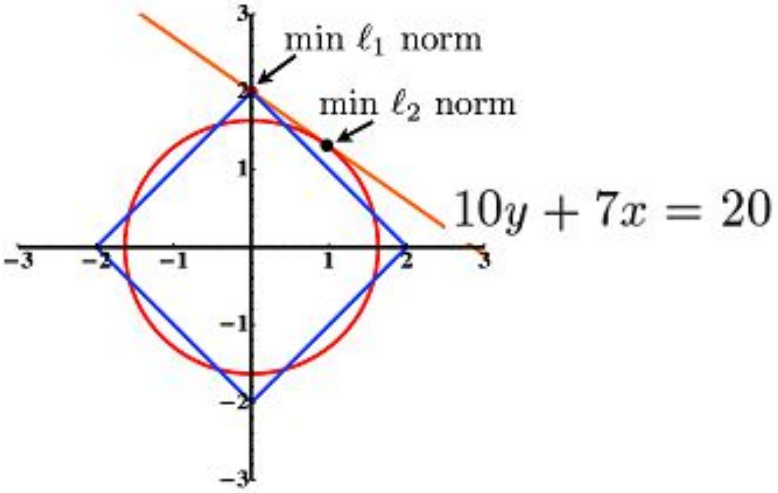
$d_{j,l}$ -- value of feature l related to material j (e.g., $|r_s(A_j) - r_p(B_j)|$) (a basis function in materials space)

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?

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



$\|c\|_1 = \sum_l |c_l|$ -- LASSO (Least Absolute Shrinkage and Selection Operator) → 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

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

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

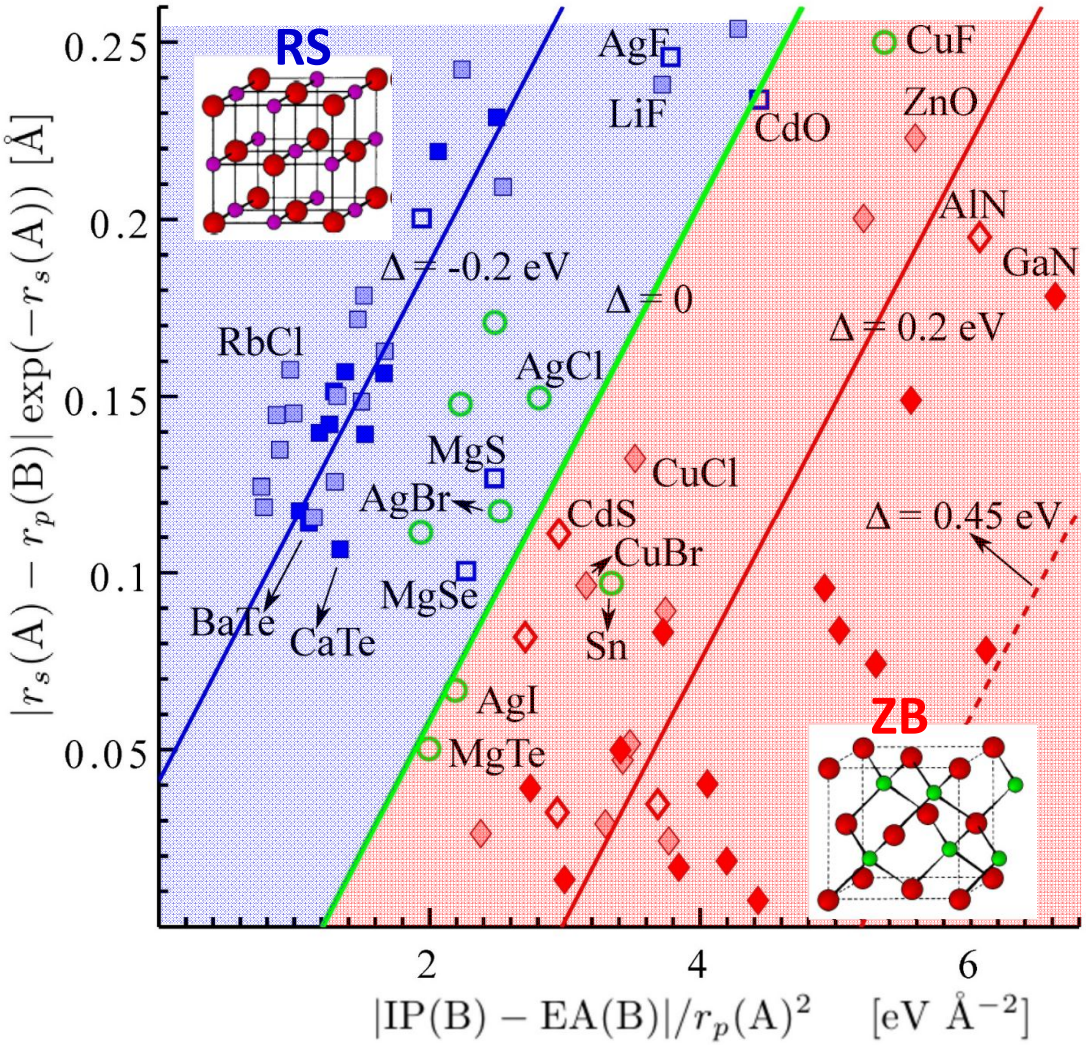
$$\Delta E = 0.113\frac{\mathrm{EA(B)}-\mathrm{IP(B)}}{r_p(\mathrm{A})^2} + 1.542\frac{|r_s(\mathrm{A})-r_p(\mathrm{B})|}{\exp(r_s(\mathrm{A}))} - 0.137 \qquad \text{2D}$$

$$\Delta E = 0.108\frac{\mathrm{EA(B)}-\mathrm{IP(B)}}{r_p(\mathrm{A})^2} + 1.790\frac{|r_s(\mathrm{A})-r_p(\mathrm{B})|}{\exp(r_s(\mathrm{A}))} + \qquad \text{3D}$$

$$+ 3.766\frac{|r_p(\mathrm{B})-r_s(\mathrm{B})|}{\exp(r_d(\mathrm{A}))} - 0.0267$$

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

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



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

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

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

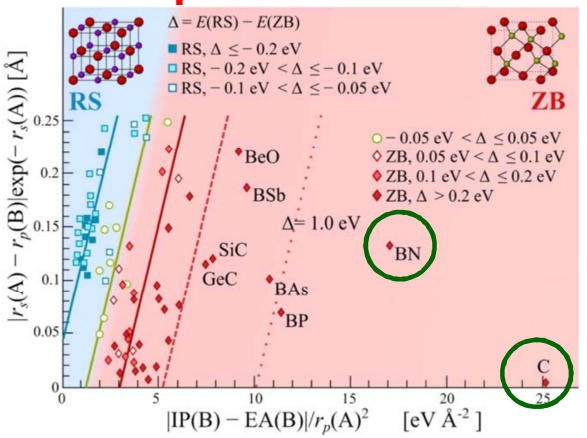
L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, and M. Scheffler,
Phys. Rev. Lett. **114**, 105503 (2015).

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:

	ΔE(LDA)	ΔE(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:

	ΔE(LDA)	ΔE(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	$1*10^{-4}$	$3*10^{-3}$	0.12	0.08	0.07	0.05
MaxAE	$8*10^{-4}$	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	$1*10^{-4}$	$3*10^{-3}$	0.12	0.08	0.07	0.05
MaxAE	$8*10^{-4}$	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

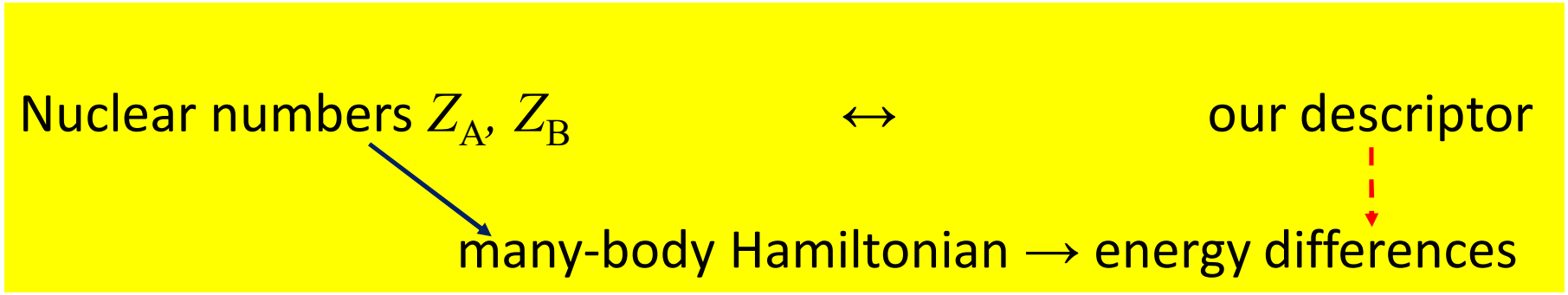
Mean absolute error (MAE), and maximum absolute error (MaxAE), in eV, (first two lines) and for a leave-10%-out cross validation (CV), averaged over 150 random selections of the training set (last two lines). For (Z_A*, Z_B*), each atom is identified by a string of three random numbers.

Descriptor	Z_A, Z_B	Z_A*, Z_B*	1D	2D	3D	5D
MAE	$1*10^{-4}$	$3*10^{-3}$	0.12	0.08	0.07	0.05
MaxAE	$8*10^{-4}$	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)

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

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

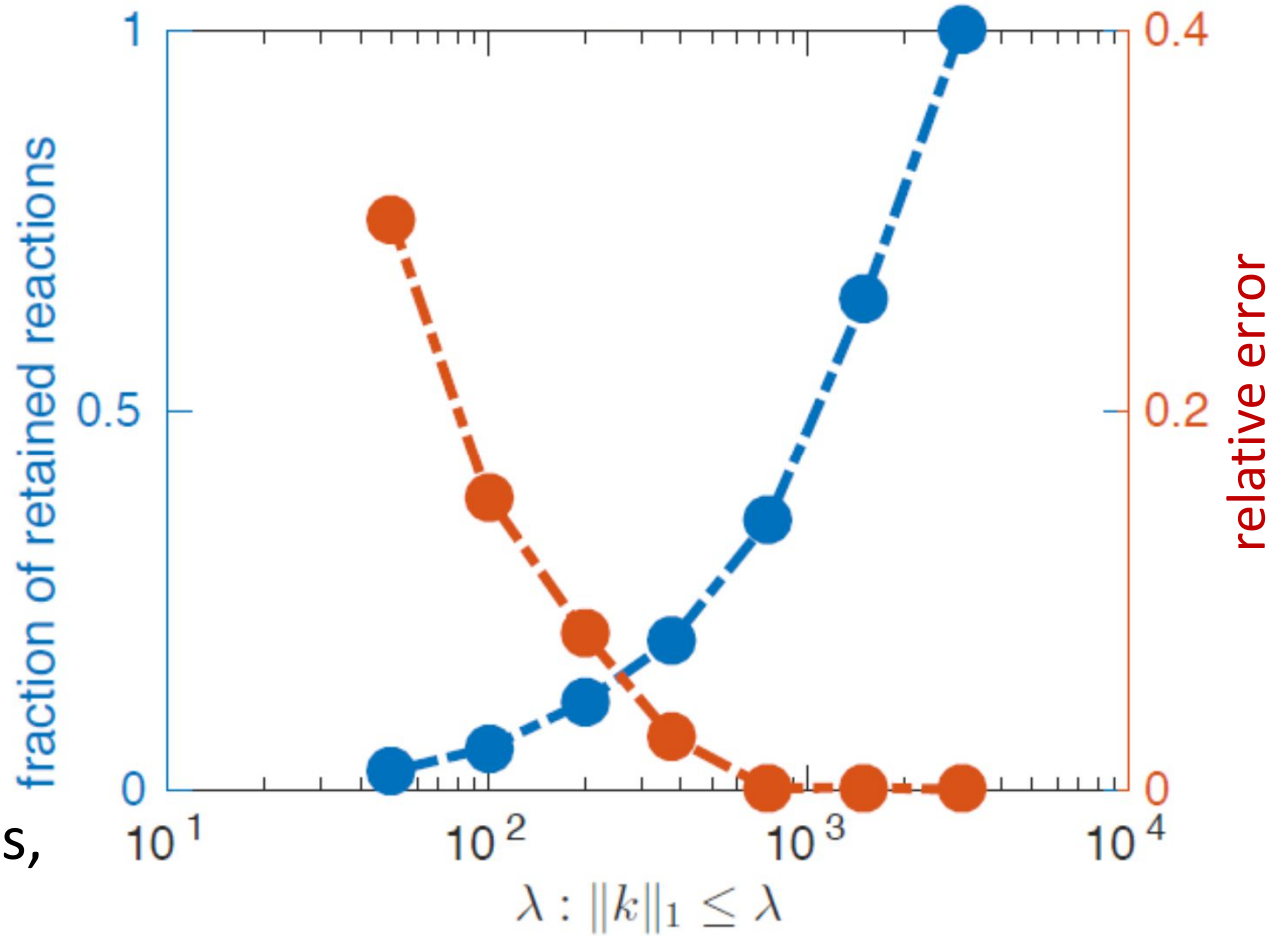
CH_4 chemical decomposition under shock-compression conditions (high *T* and *p*)

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

Methane at T = 3,300 K, p = 40.53 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.

$$\min_{\hat{k}} \left\| A\hat{\boldsymbol{k}} - b \right\|_2$$
$$\text{subject to } \hat{\boldsymbol{k}} \geq 0, \left\| \hat{\boldsymbol{k}} \right\|_1 \leq \lambda$$

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



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

$$\boldsymbol{F}_a = -\Phi_a - \Phi_{ab}\boldsymbol{u}_b - \frac{1}{2}\Phi_{abc}\boldsymbol{u}_b\boldsymbol{u}_c - \cdots$$

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)

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

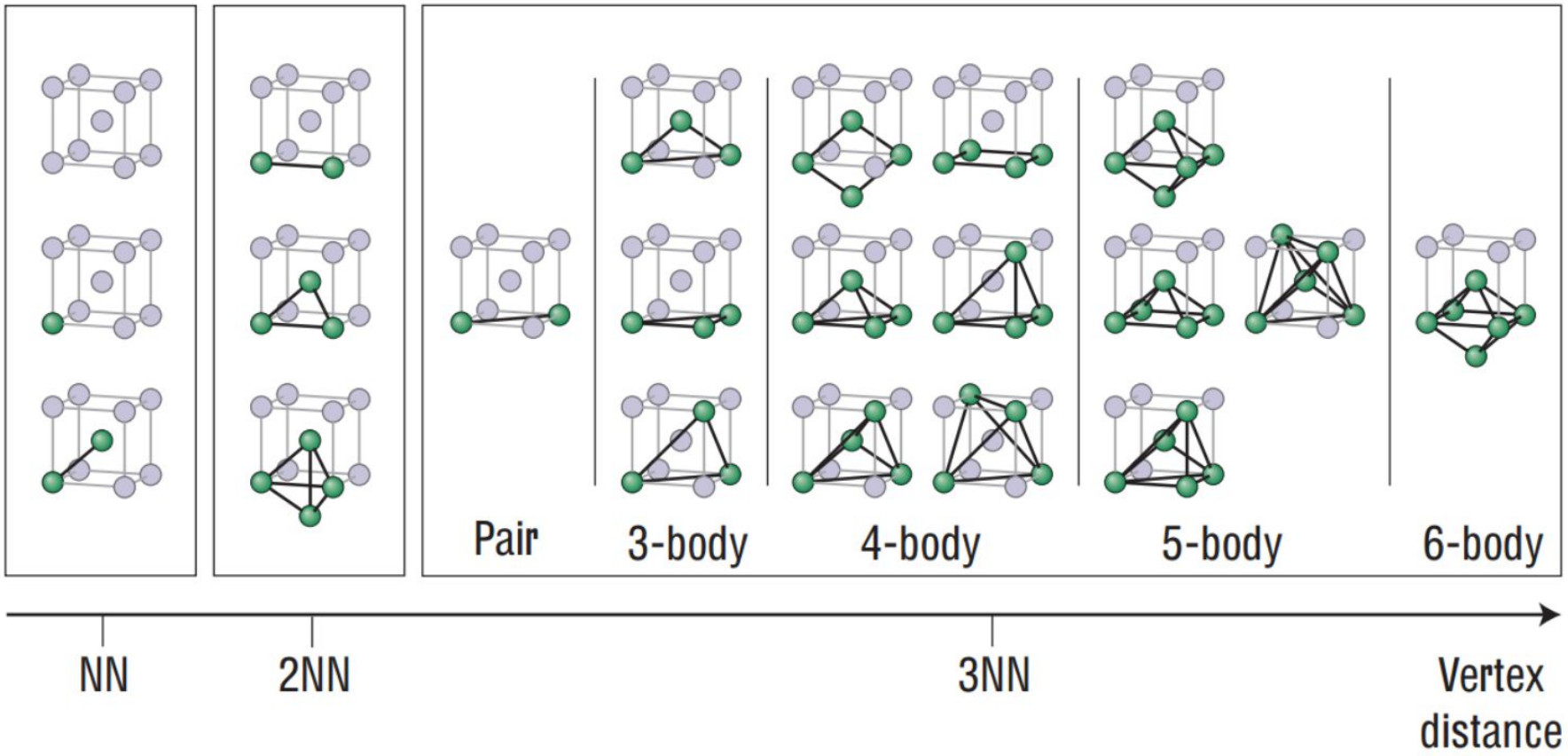
$$A_{aJ} = \begin{bmatrix} -1 & u_b^1 & -\frac{1}{2}u_b^1u_c^1 & \cdots \\ \vdots & \vdots & \vdots & \\ -1 & u_b^L & -\frac{1}{2}u_b^Lu_c^L & \cdots \end{bmatrix}$$

→ predictive model for anharmonic lattice dynamics

F. Zhou, W. Nielson, Y. Xia, and Vidvuds Ozoliņš, Phys. Rev. Lett. 113, 185501 (2014)

Compressive Sensing for Cluster Expansion

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



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 → convex problem, equivalent to the NP-hard if features are uncorrelated → not the case when many features are generated → Sure Independence Screening plus Selection Operator (SISSO)**

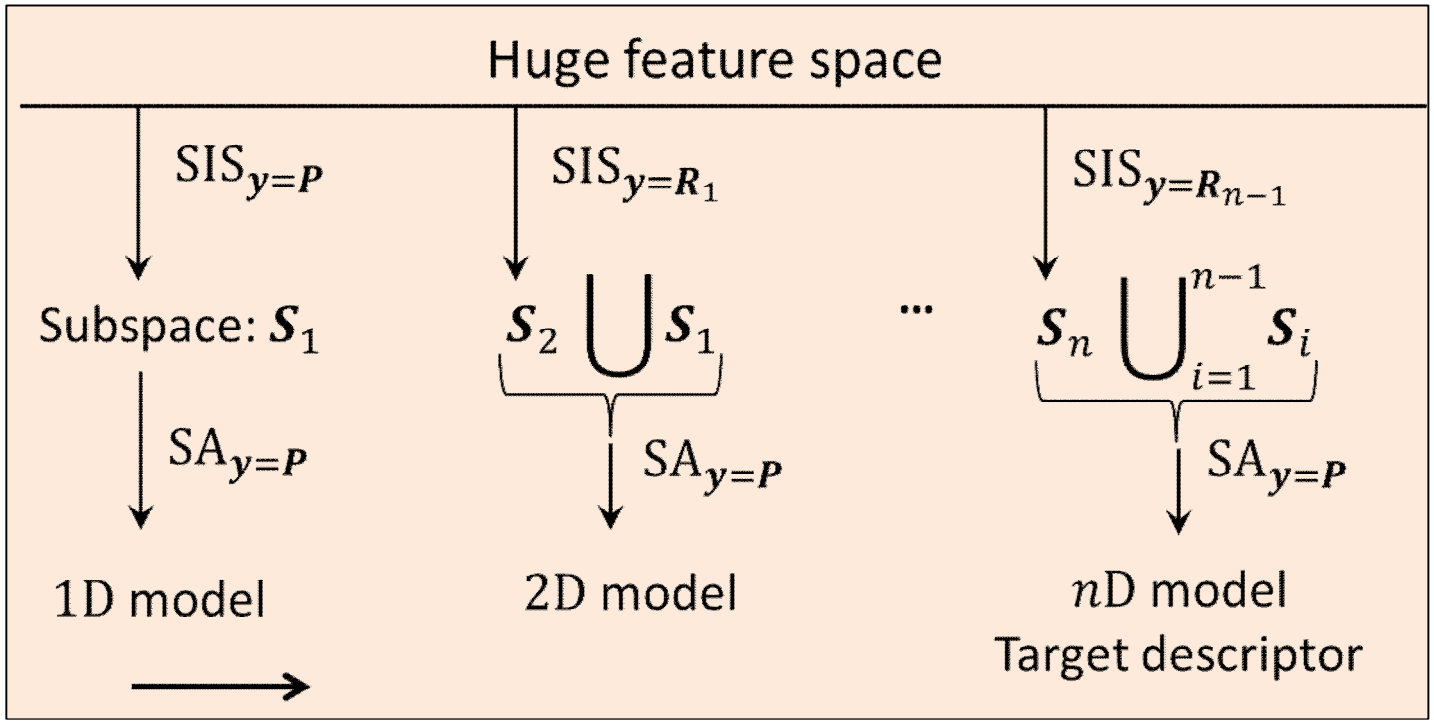
1. Systematically construct a huge feature space (10^{11}) from primary features: $\hat{R} = \{+, -, \cdot, ^{-1}, ^2, ^3, \sqrt{\ }, exp, log, |-|\}$(use physically meaningful combinations!)
2. Select top ranked features using *Sure Independence Screening (SIS)*[1] (correlation learning). Select n features corresponding to the n largest projection on the target property, i.e. largest components of the vector ($\boldsymbol{D}^T \boldsymbol{y}$)

   $\boldsymbol{y}$: vector with the target property (e.g., rock salt-zincblende energy differences; 82 elements)

   $\boldsymbol{D}$: 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

R. Ouyang, *et al.*, Physical Review Materials 2, 083802 (2018)

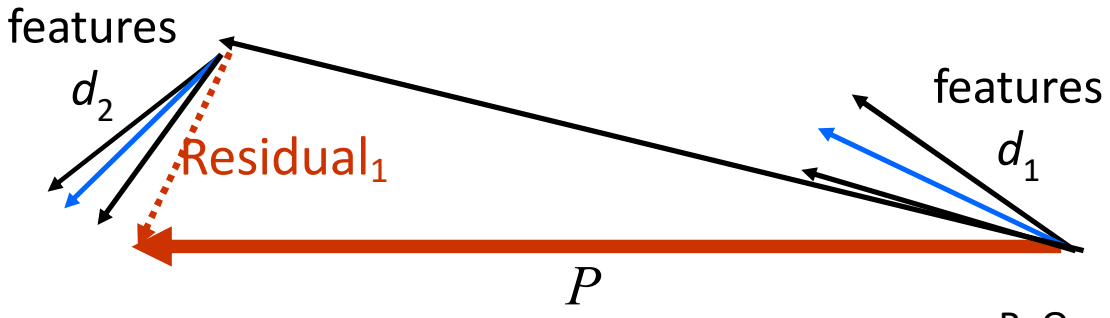
SISSO: Iterative residual fitting



$\boldsymbol{y}$: response vector

$\boldsymbol{P}$: target material property

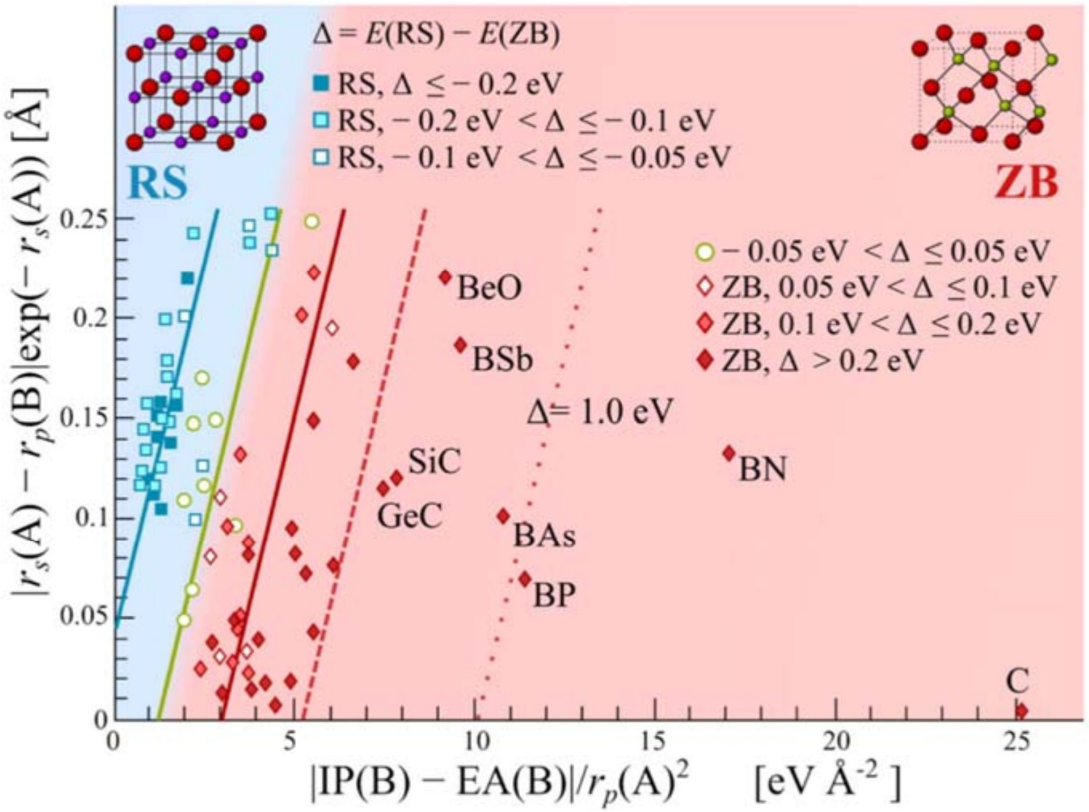
Residual: $R = P - \sum_i c_i d_i$



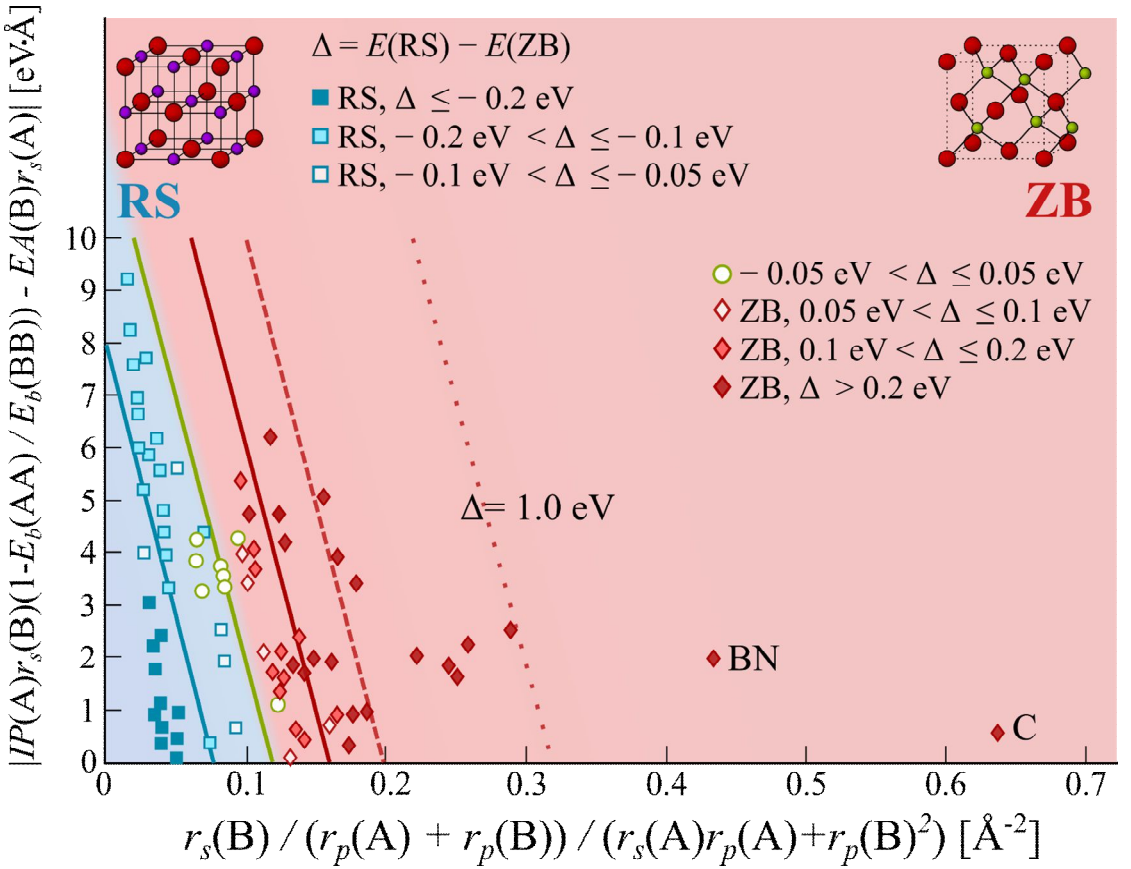
R. Ouyang, *et al.*, Physical Review Materials 2, 083802 (2018)

SISSO: Performance

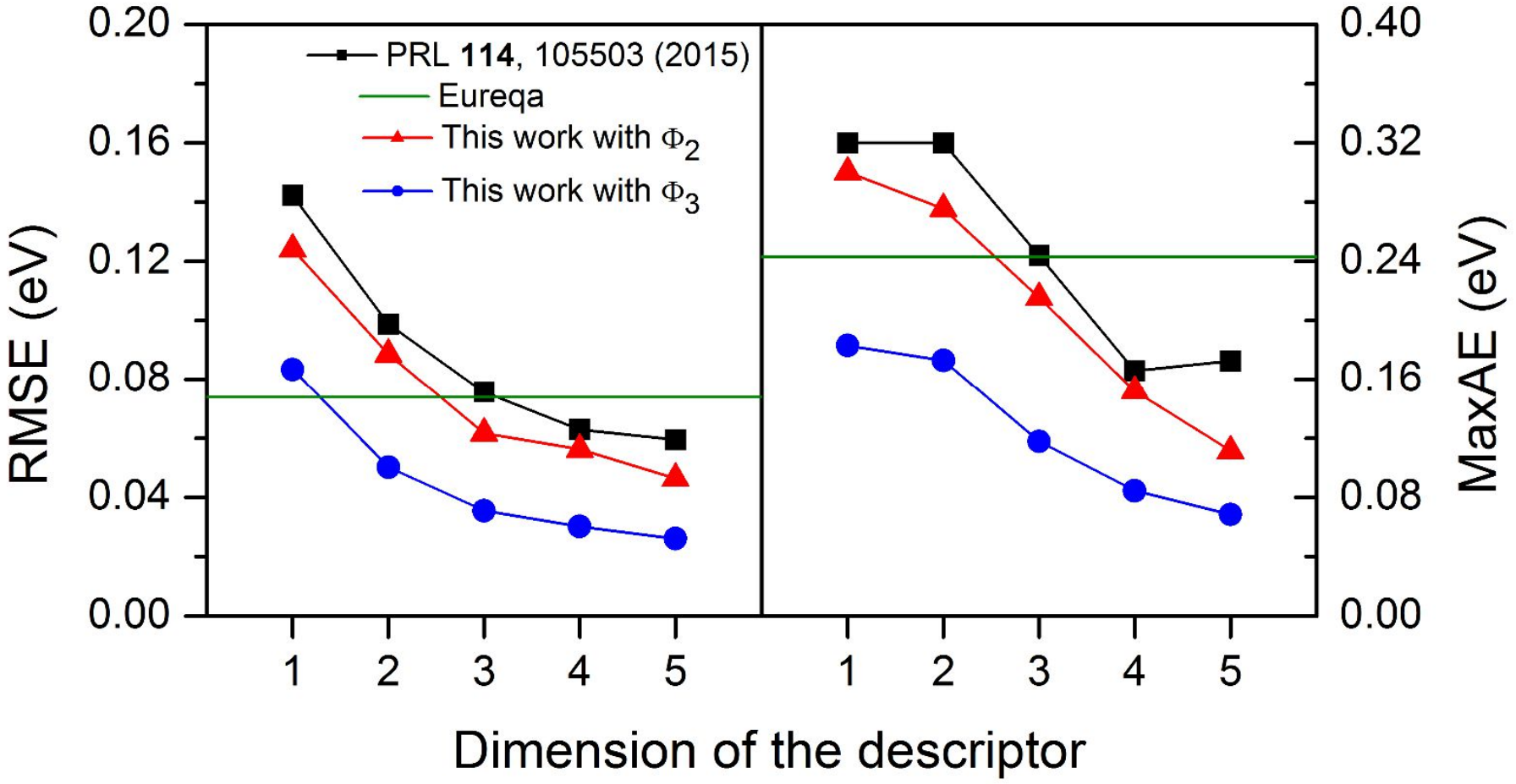
LASSO(+l_0)



SISSO



SISSO: Performance



SISSO: Multitask and categorical

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

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

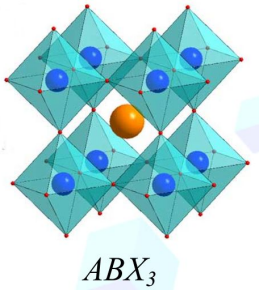
Categorical (can be also multitask): Property - material belongs to a given class (yes/no)

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

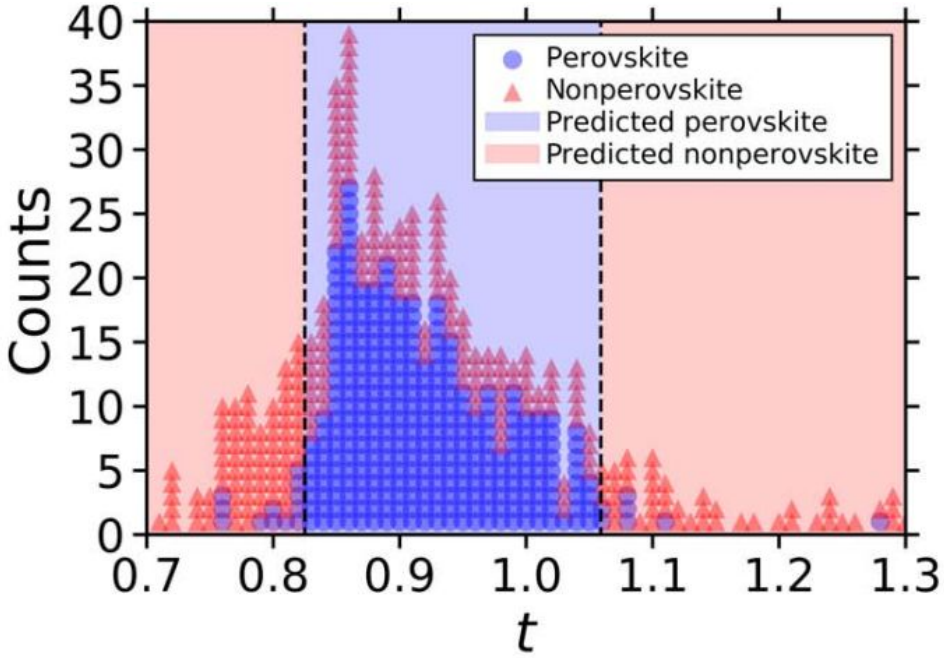
number of data in the overlap region between domains of different classes in d-space

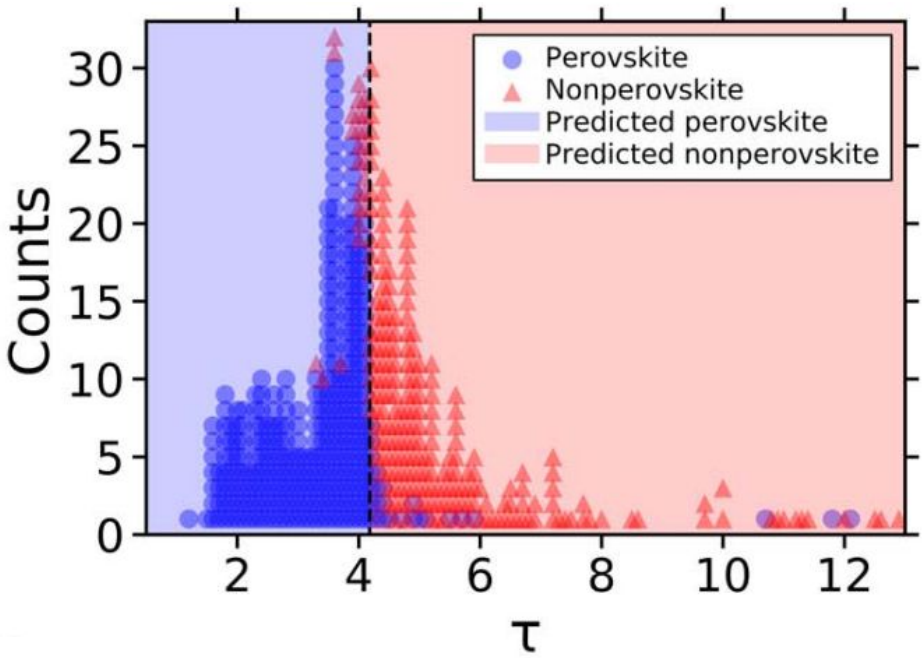
R. Ouyang, *et al*., J. Phys.: Mater. 2, 024002 (2019)

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%

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

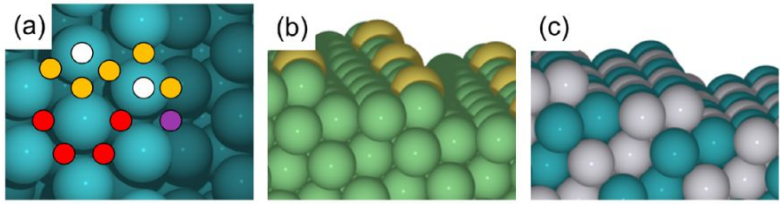
oxidation state

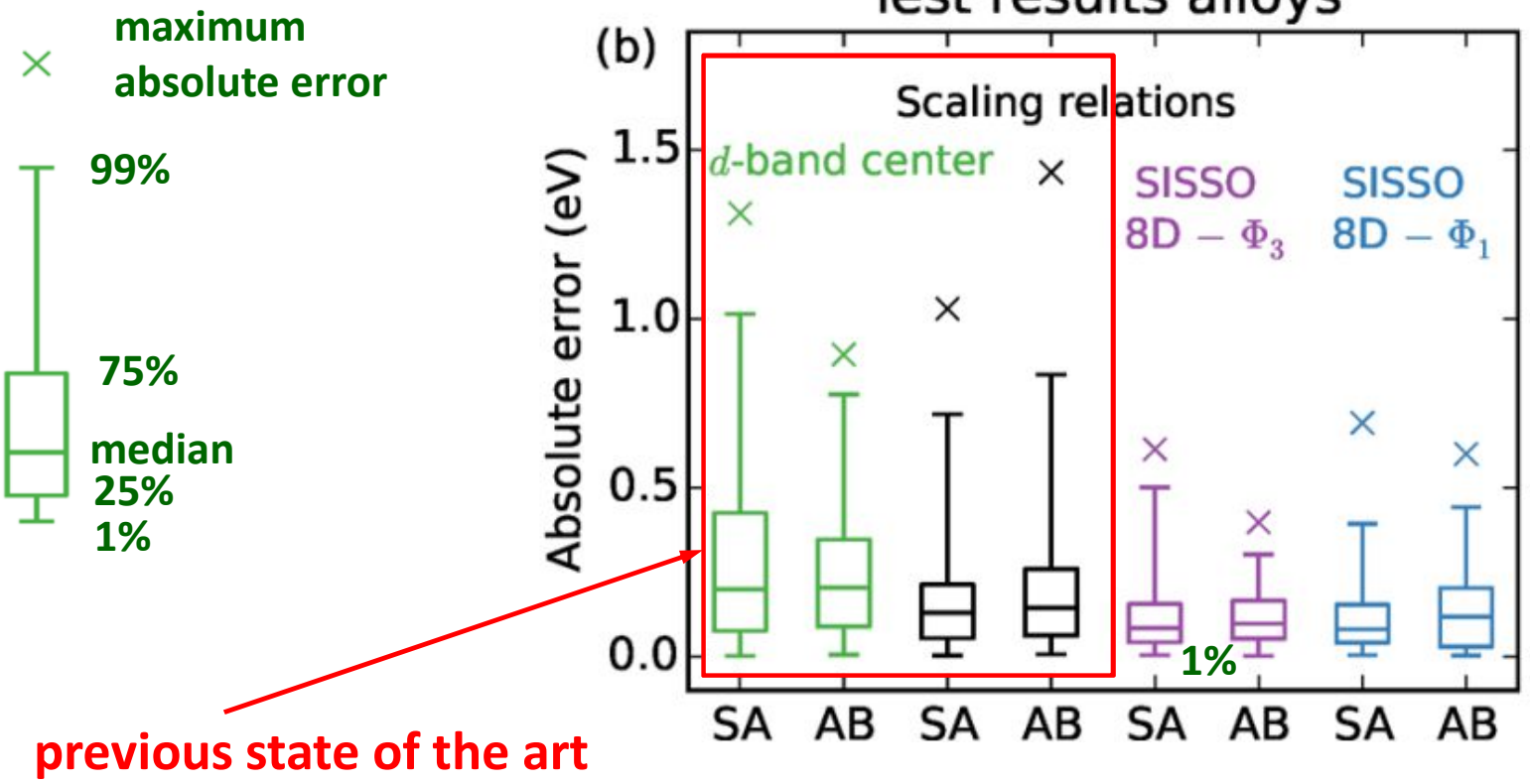
C. Bartel *et al.*, Sci. Adv. *5, eaav0693* (2019)

SISSO: Examples

- Adsorption of molecules on metal surfaces

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

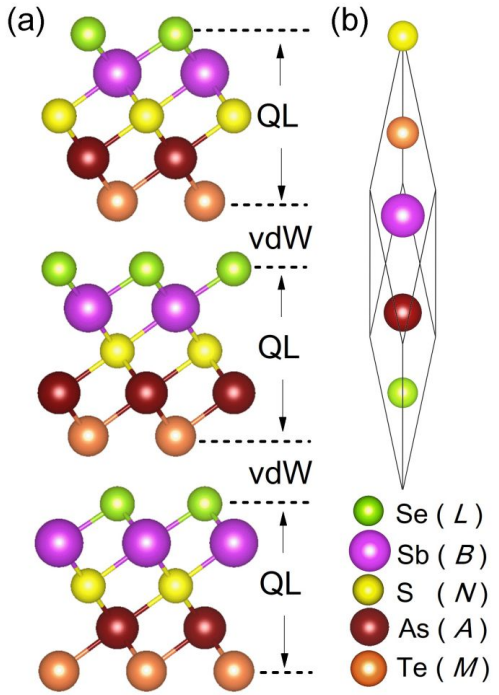


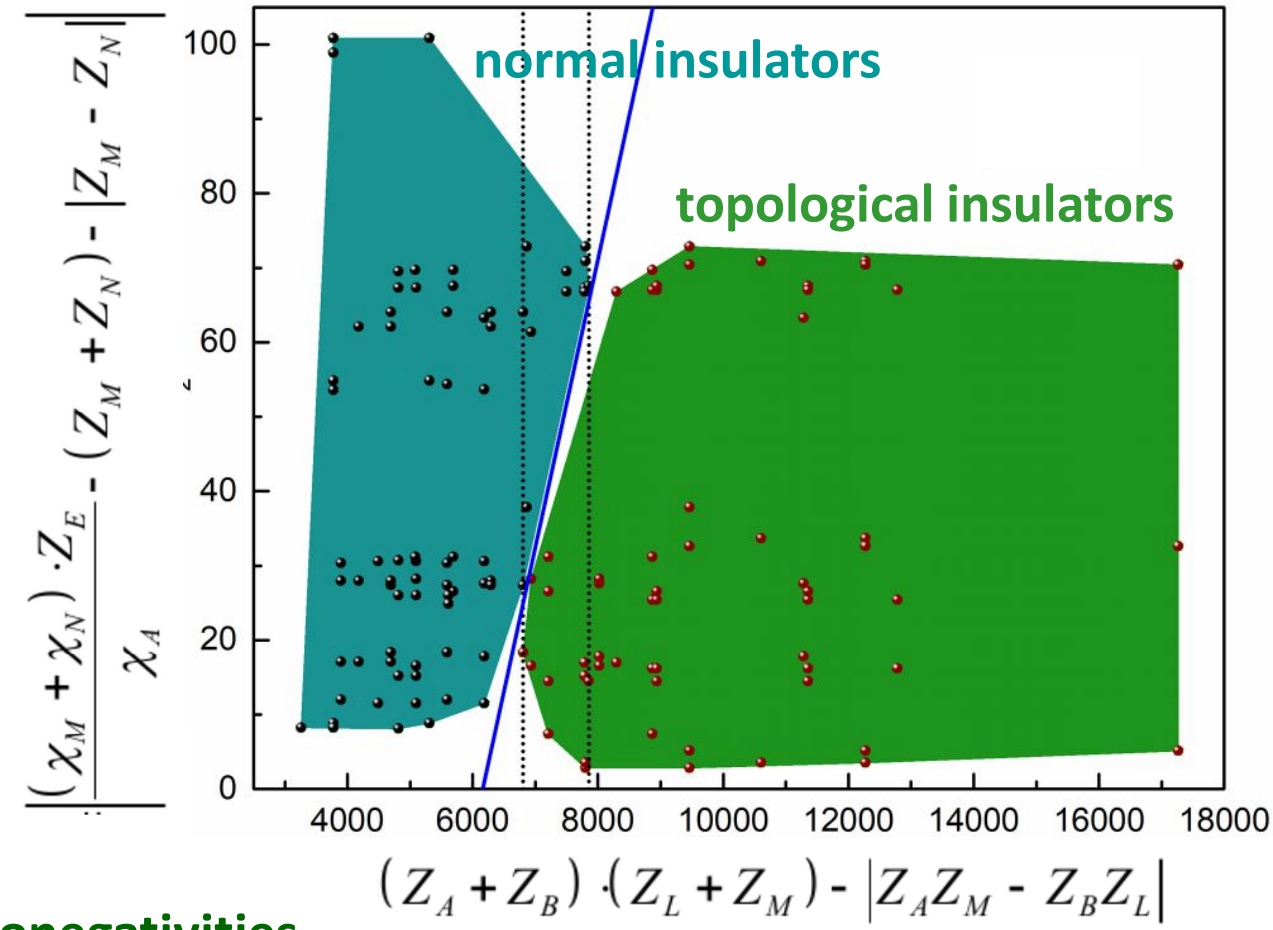


M. Andersen *et al*., ACS Catal. 9, 2752 (2019)

SISSO: Examples

- **Design of topological insulators (materials for spintronics, catalysis, thermoelectricity)**

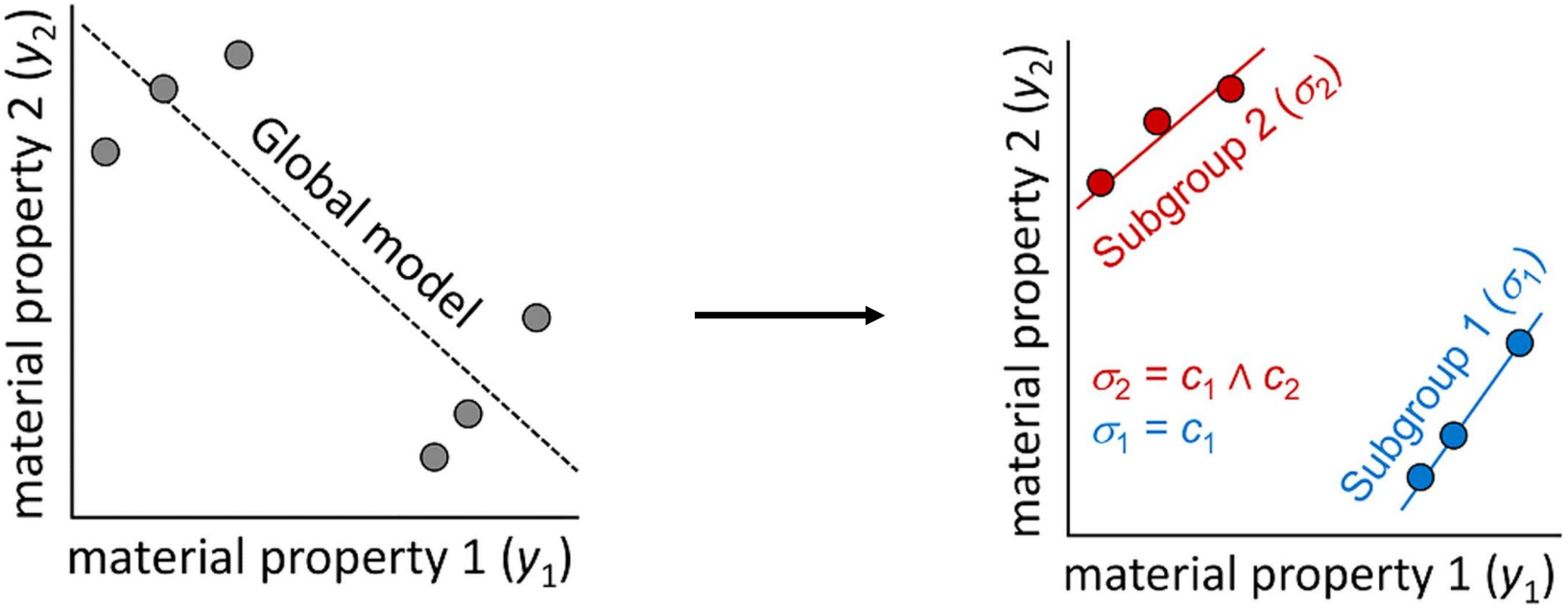




Z - atomic numbers, χ - electronegativities

G. Cao *et al.*, arXiv:1808.04733

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*

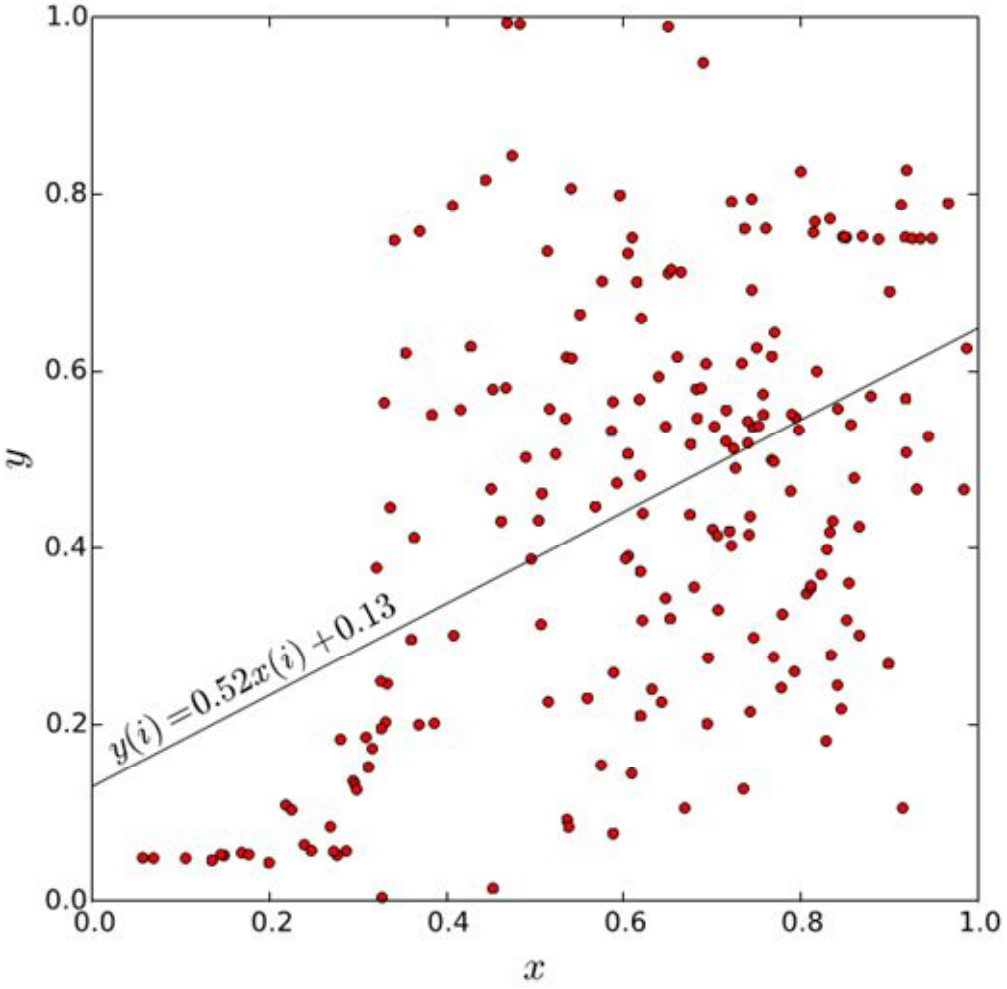
$$f = N_{subgroup}/N_{all} \times |mean_{subgroup} - mean_{all}| \times (1 - variance_{subgroup}/variance_{all})$$

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

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

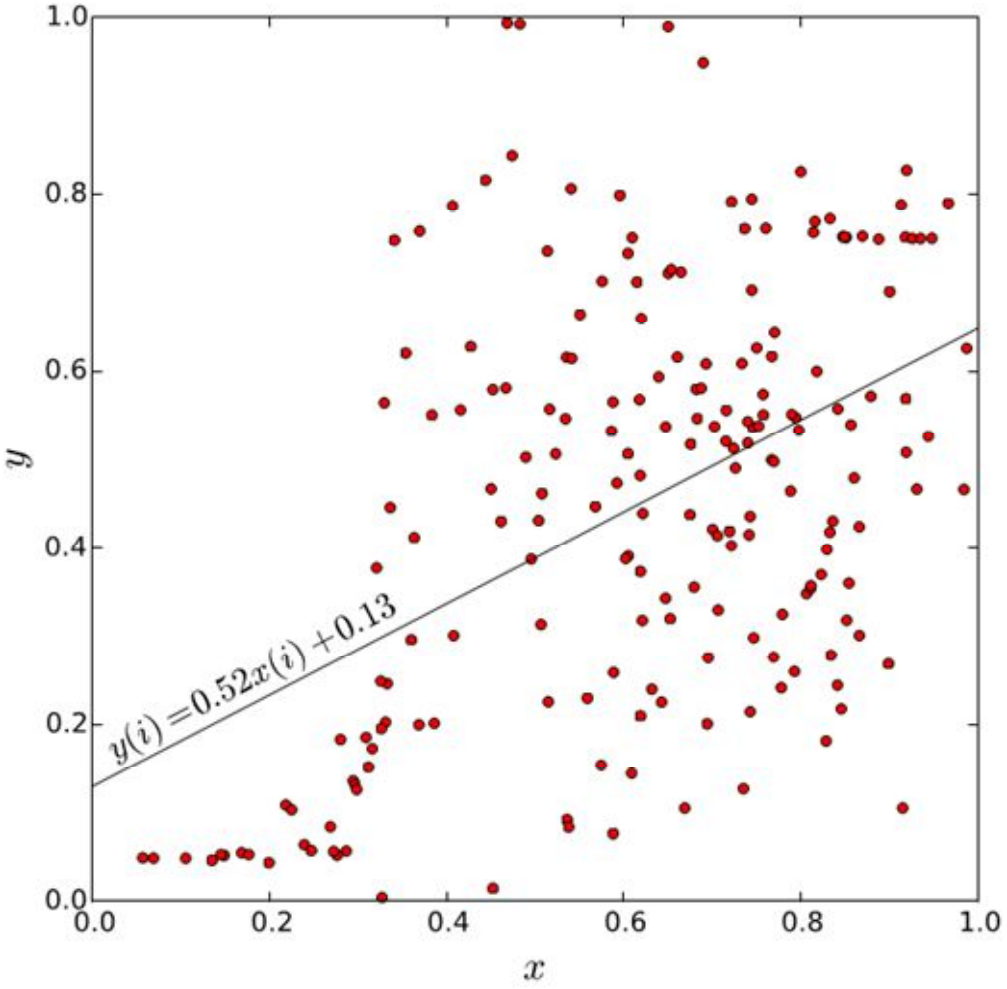
W. Klösgen, Advances in Knowledge Discovery and Data Mining. Palo Alto, CA: AAAI Press; 1996, 249

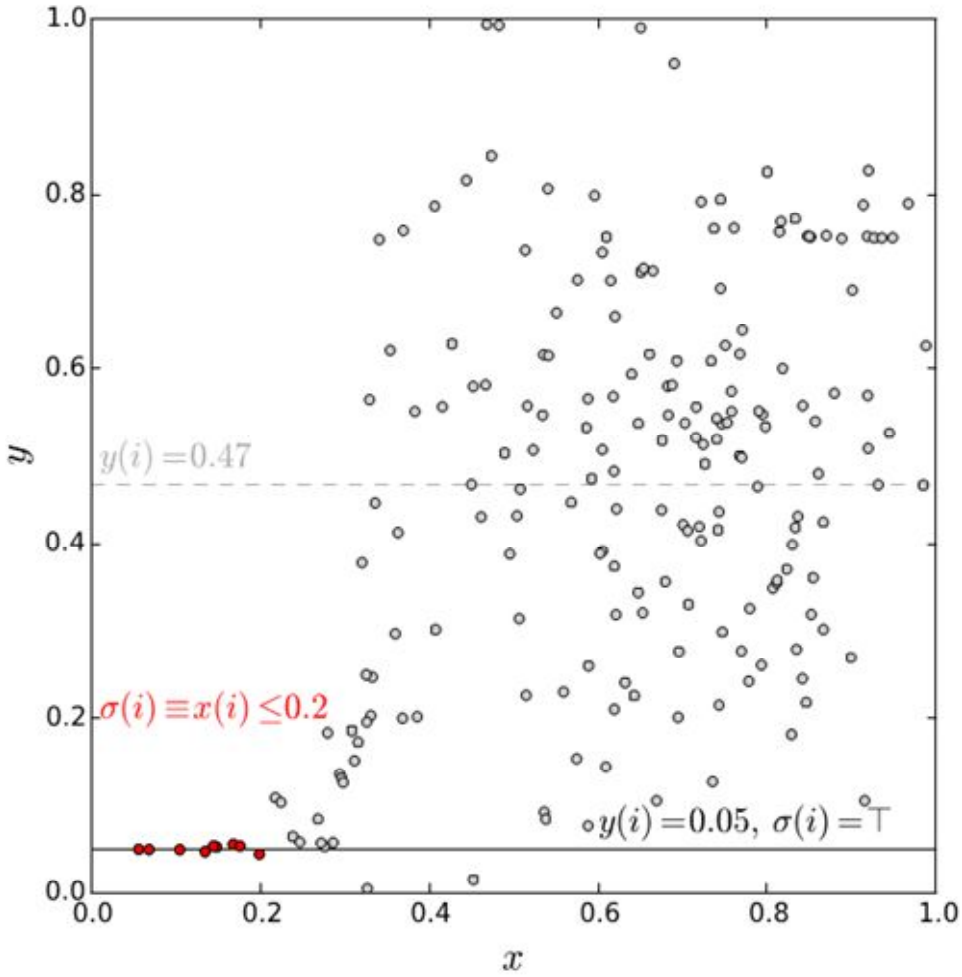
Data mining: Subgroup discovery



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

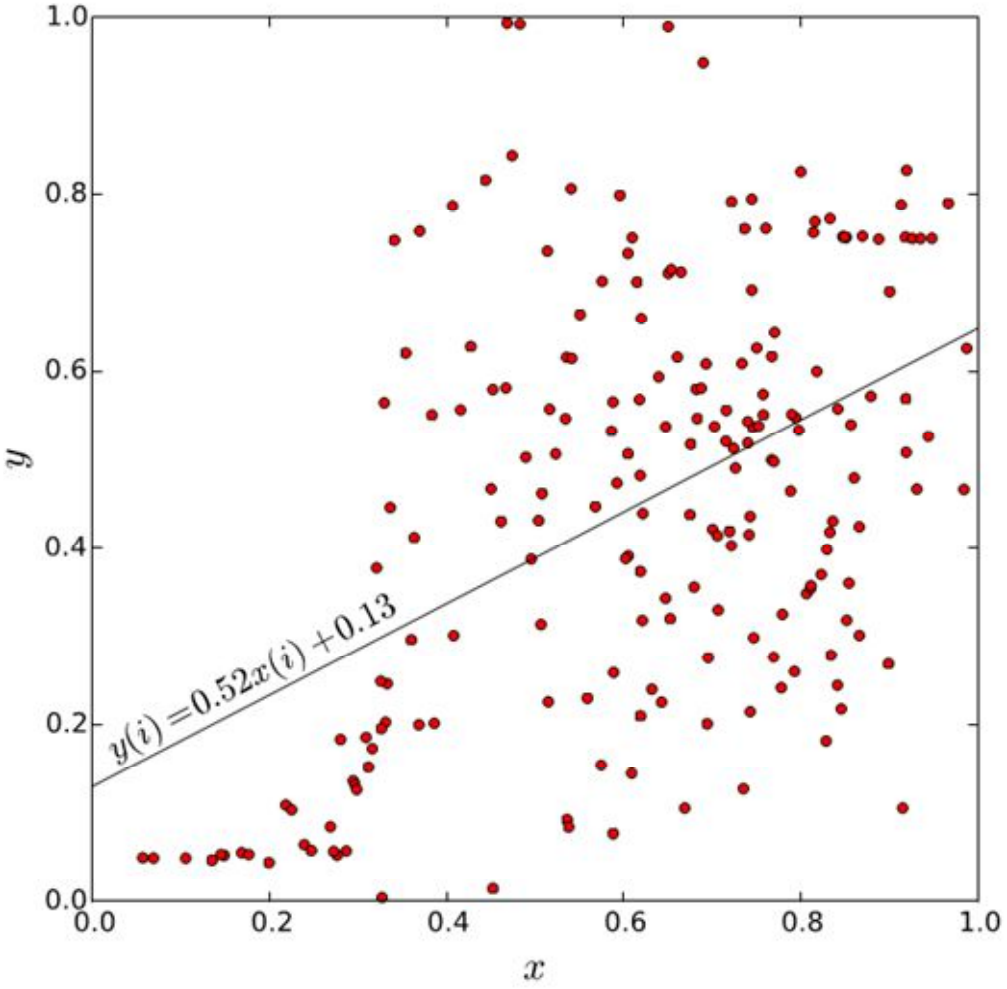
Data mining: Subgroup discovery

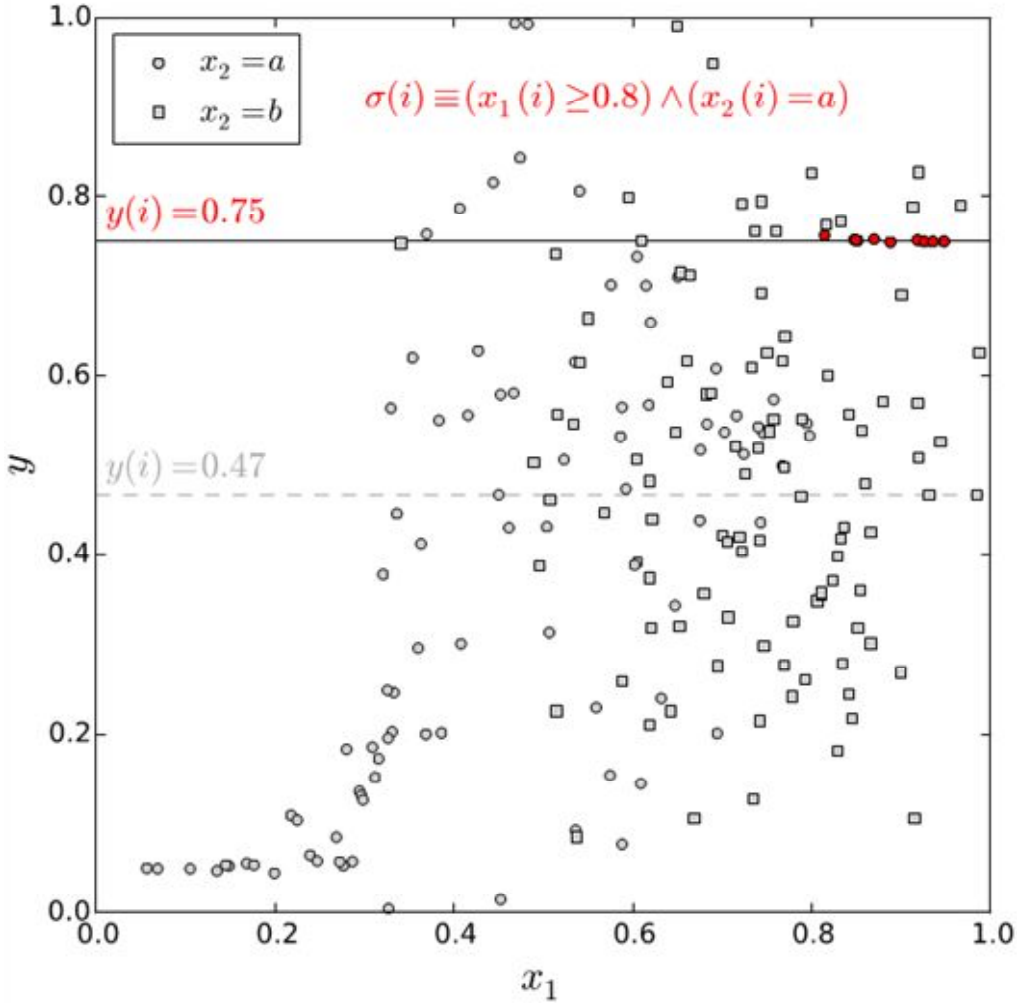




M. Boley *et al*., Data Min. Knowl. Disc. 31, 1391 (2017); B. Goldsmith *et al*., New J. Phys. 19, 013031 (2017)

Data mining: Subgroup discovery



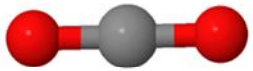


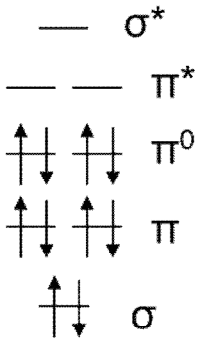
M. Boley *et al*., Data Min. Knowl. Disc. 31, 1391 (2017); B. Goldsmith *et al*., New J. Phys. 19, 013031 (2017)

Subgroup discovery: CO_2 activation by adsorption

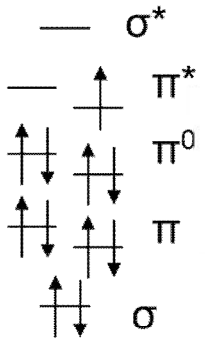
CO_2

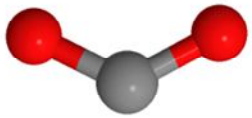
C-O bond elongation, O-C-O bending angle → indicators of activation



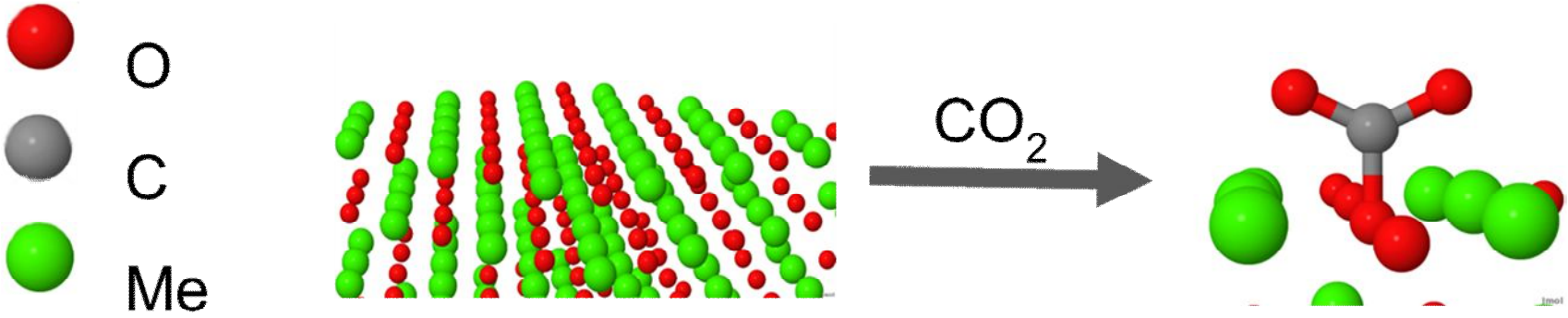


charge transfer





Subgroup discovery: CO_2 activation by adsorption



dry reforming of methane:
$CO_2 + CH_4 = 2H_2 + 2CO$

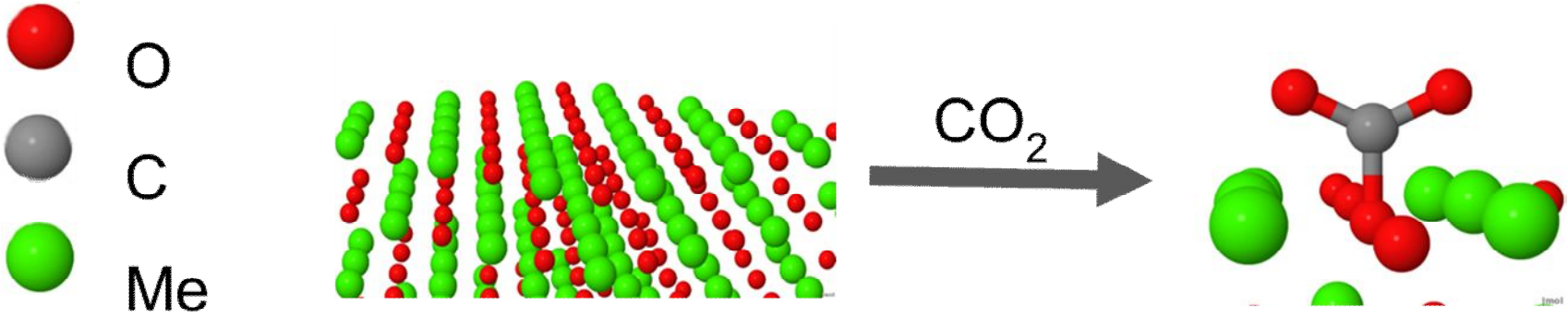
Sabatier reaction:
$CO_2 + 4H_2 = CH_4 + 2H_2O$

partial hydrogenation:
$CO_2 + 3H_2 = CH_3OH + H_2O$

Oxides:

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

Subgroup discovery: CO_2 activation by adsorption



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

Which surface properties lead to desired indicators?

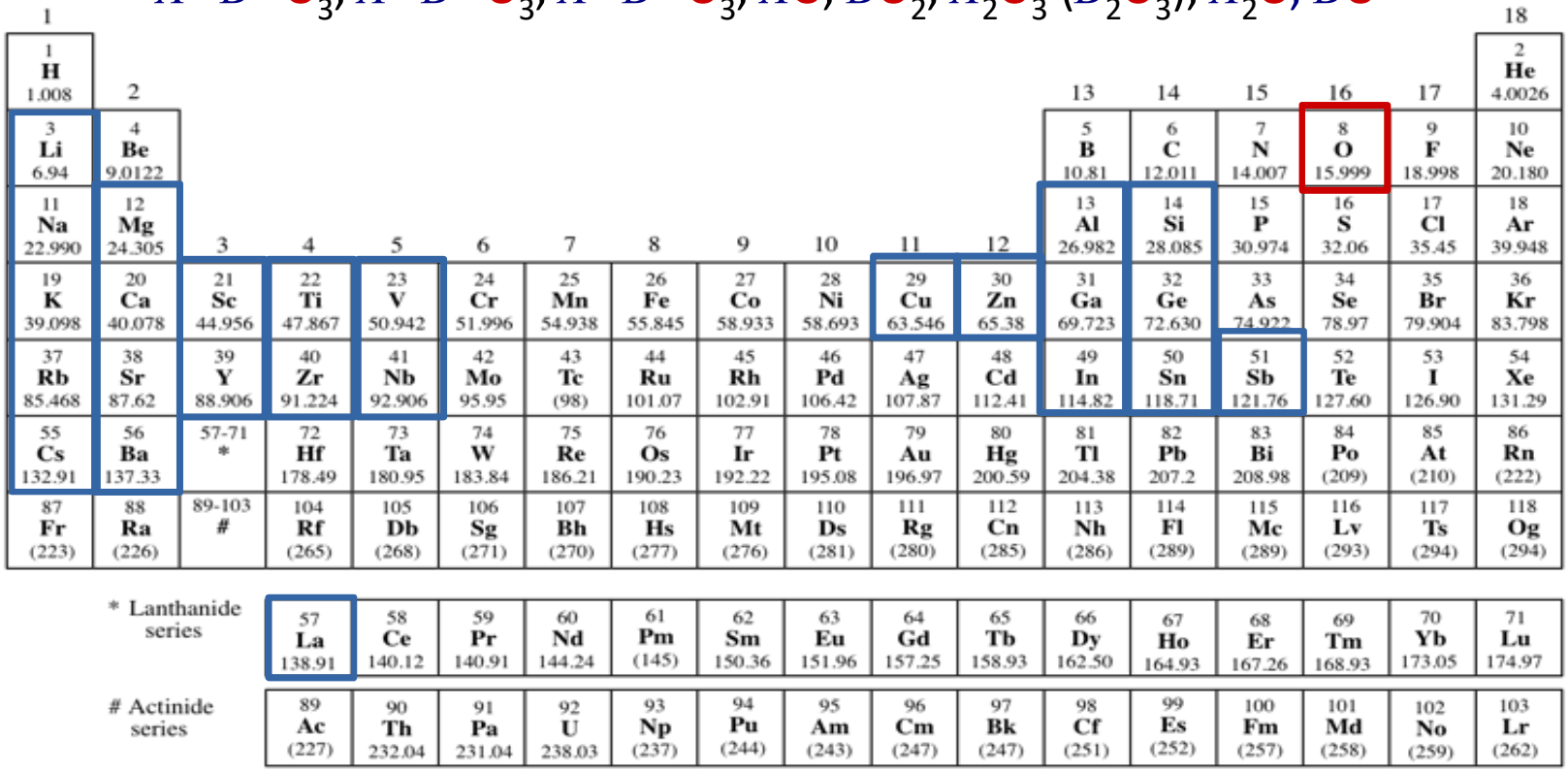
Use subgroup discovery to find materials that optimize activation indicators

$$f = N_{subgroup}/N_{all} \times (mean_{subgroup} - mean_{all}) \times (1 - variance_{subgroup}/variance_{all})$$

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

Subgroup discovery: CO_2 activation by adsorption

$A^{2+}B^{4+}O_3$, $A^{3+}B^{3+}O_3$, $A^{1+}B^{5+}O_3$, AO, BO_2, A_2O_3 (B_2O_3), A_2O, BO

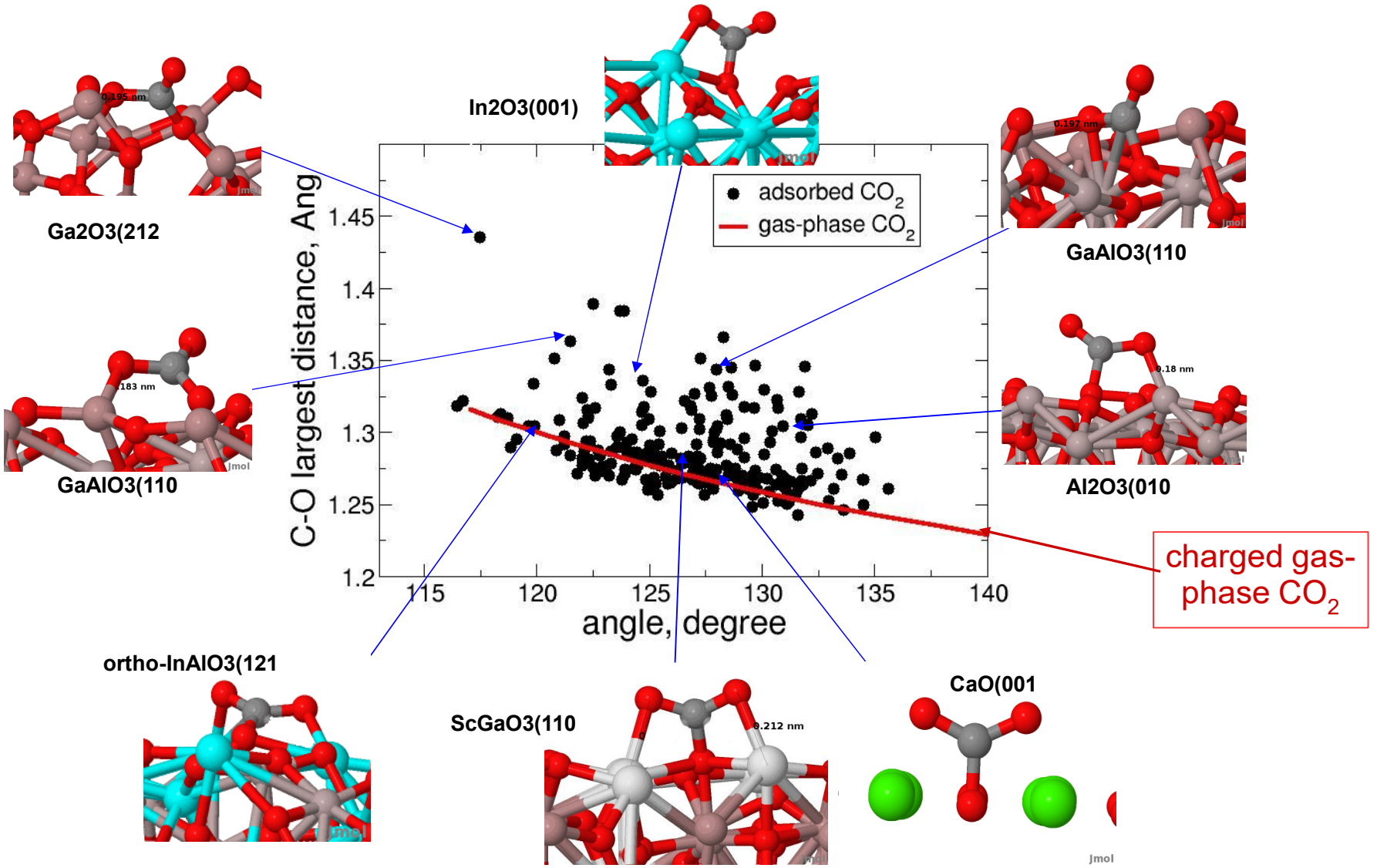


71 oxide materials

141 surfaces with Miller indexes ≤ 2

270 adsorption sites

Subgroup discovery: CO_2 activation by adsorption



Primary features

Atom:

electron affinity ionization potential electronegativity

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

Material:

work function band gap Cbm surface form. energy

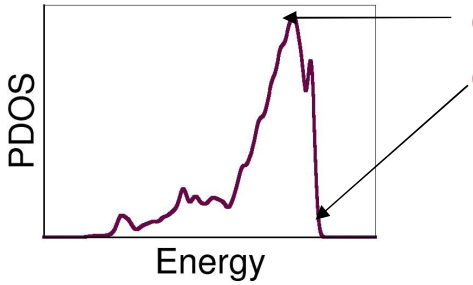
Site-specific features:

electrostatic potential Hirshfeld charge bond-valence of O

coordination number of O vdW C_6-coefficient polarizability

distances to 1st, 2nd, 3d nearest cations local-structure parameters

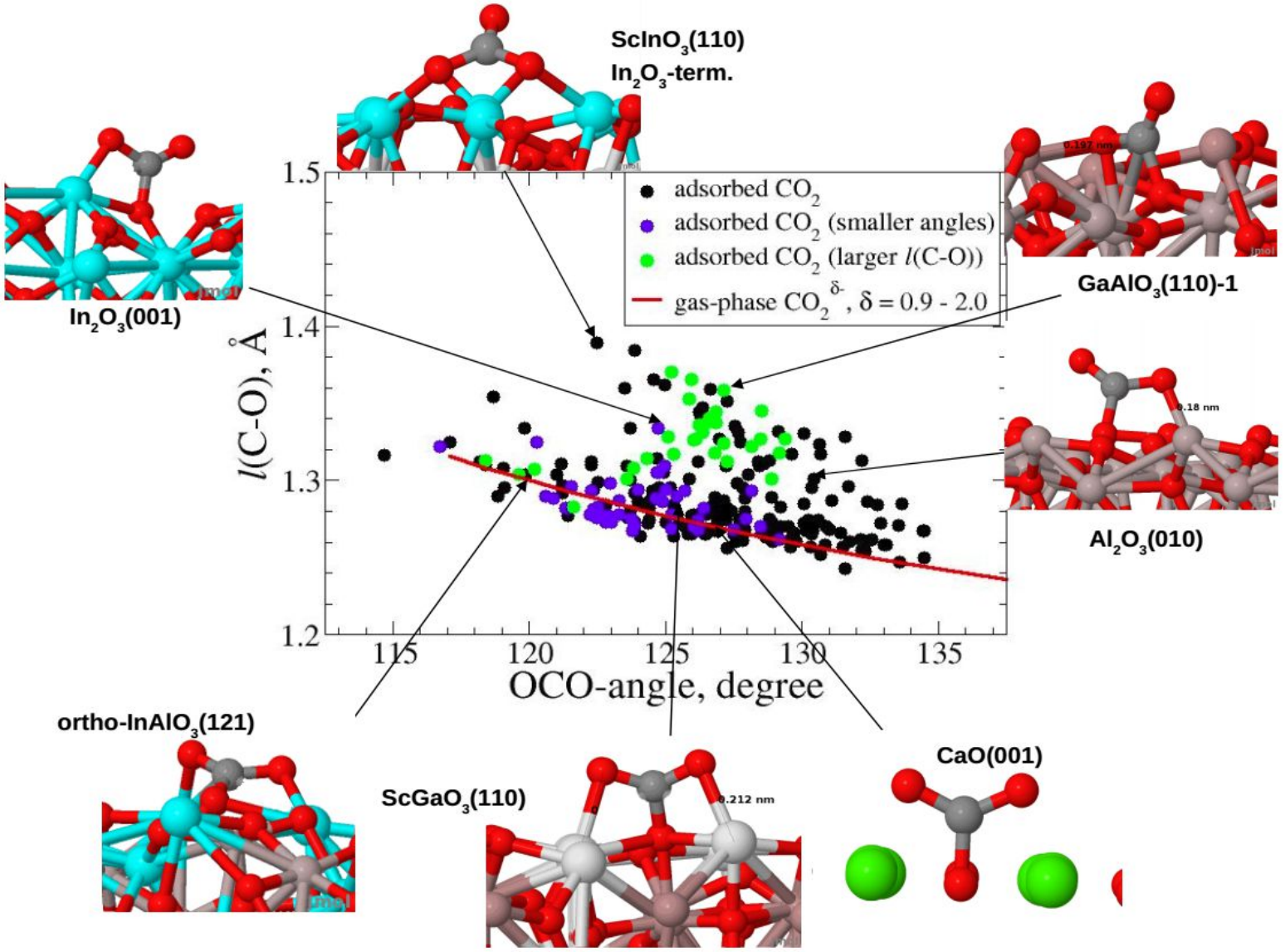
features of
O 2*p*-PDOS



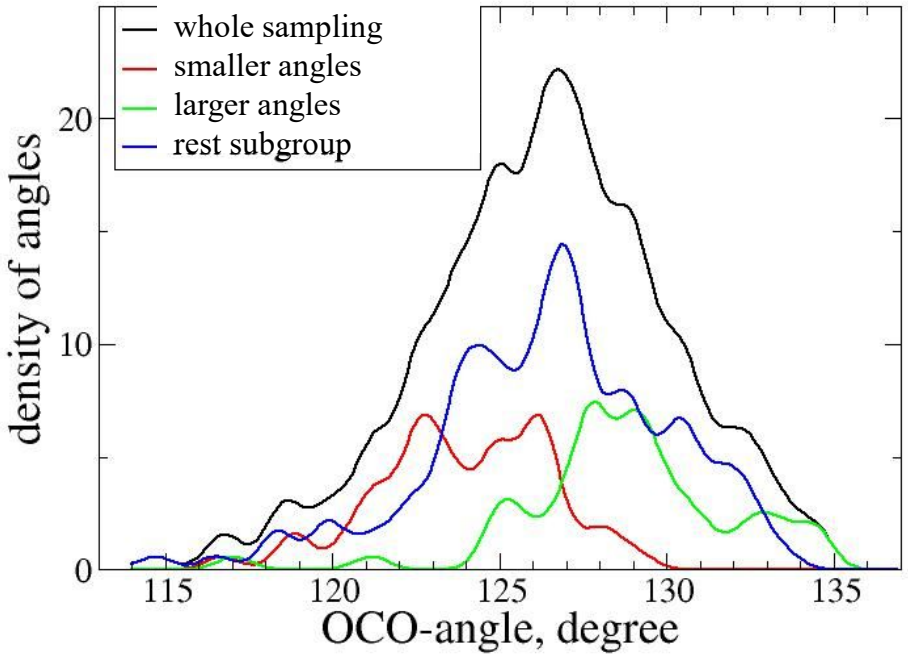
energy of maximum
energy of top

1st, 2nd, 3d, 4th moments
DOS moments: center, width, skewness, kurtosis

Subgroup discovery: Adsorbed CO2 properties

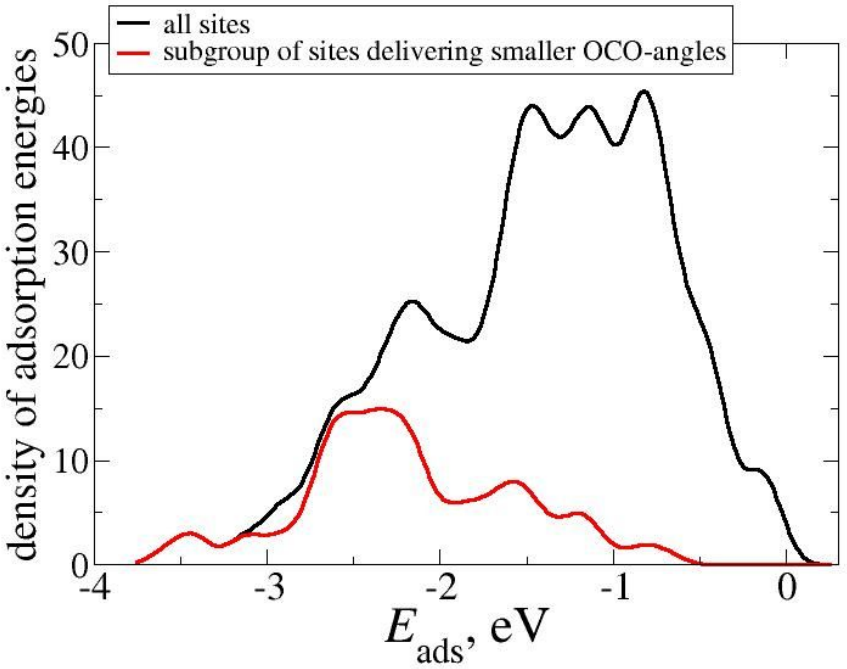


Subgroup discovery: Analysis of the OCO angle



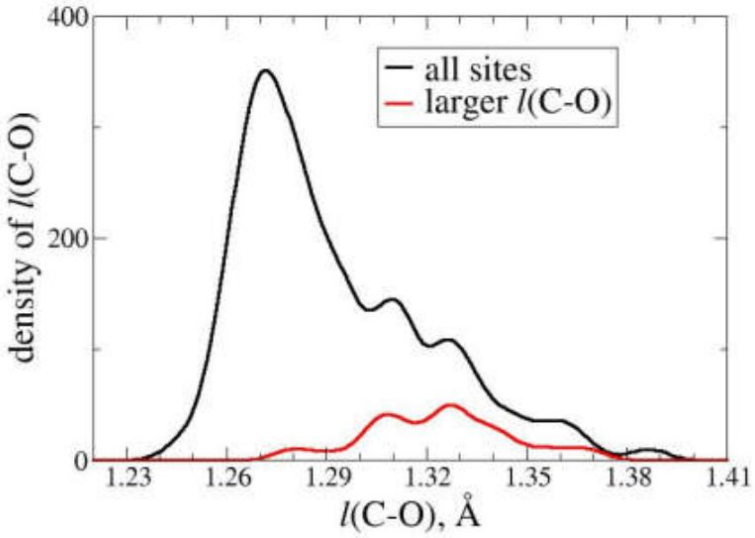
sites delivering smaller angles (59 adsorption sites):

(energy of O 2*p* 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)



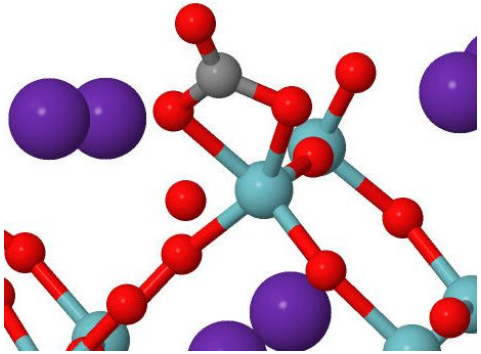
Most of the site delivering smaller OCO angles are on ionic (basic) materials

Subgroup discovery: Analysis of the C-O bond length



sites delivering larger *l*(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)



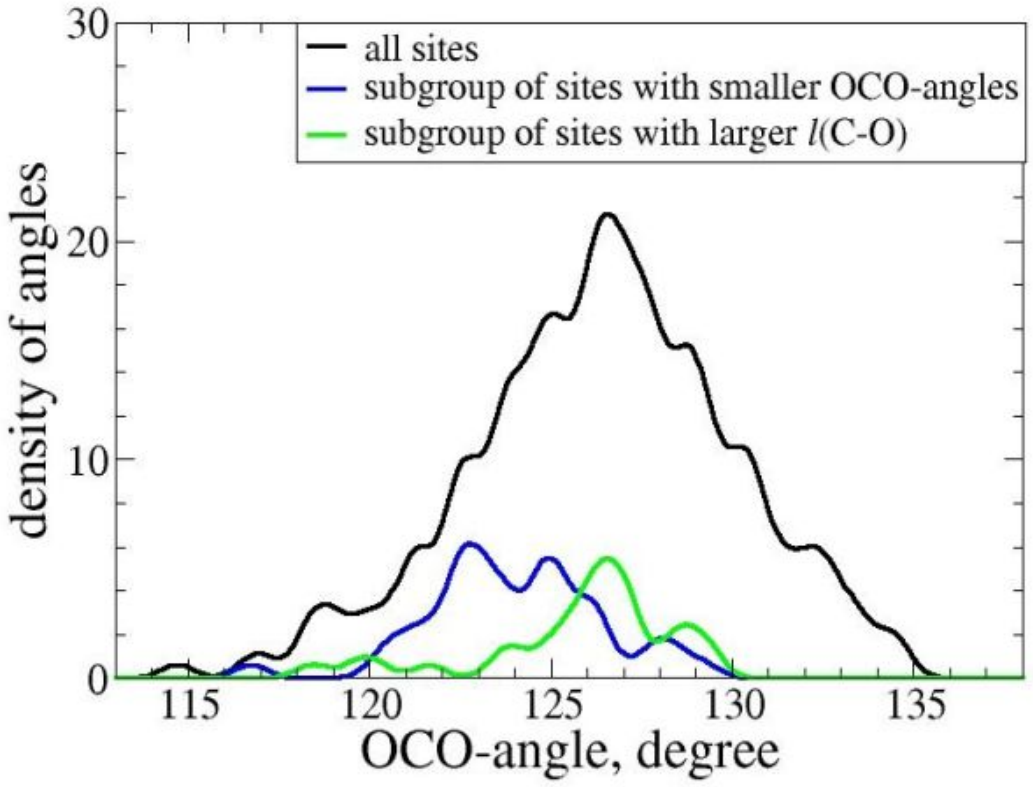
$LaGaO_3$ – cathode material in high-temperature electrochemical CO_2 reduction;

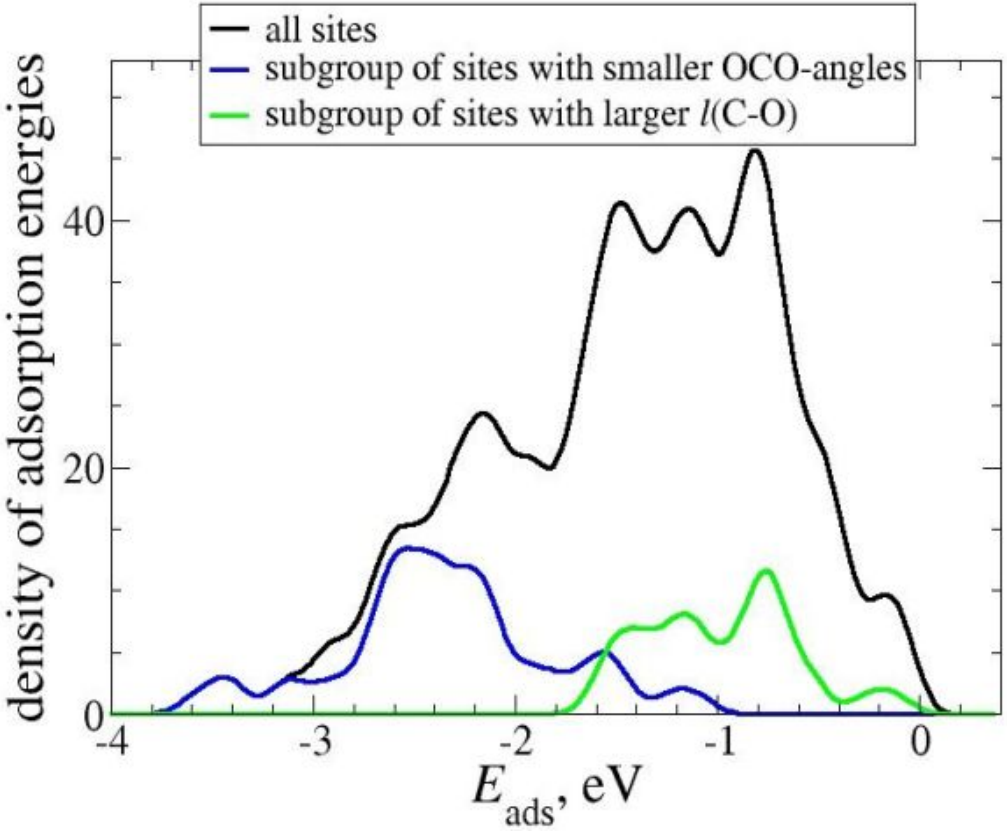
$KNbO_3$ – photocatalytic reduction of CO_2 into CH_4;

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

$NaSbO_3$ – material for CO_2 capture and storage (CCS)

Subgroup discovery: Alternative mechanisms of CO_2 activation



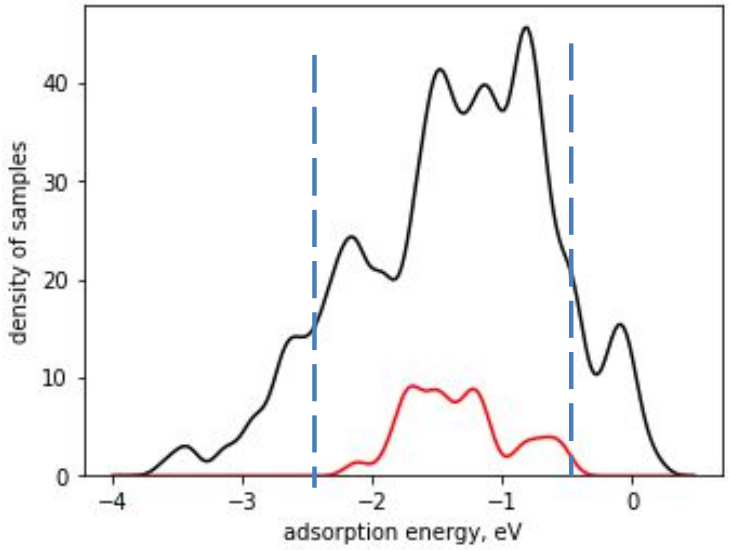


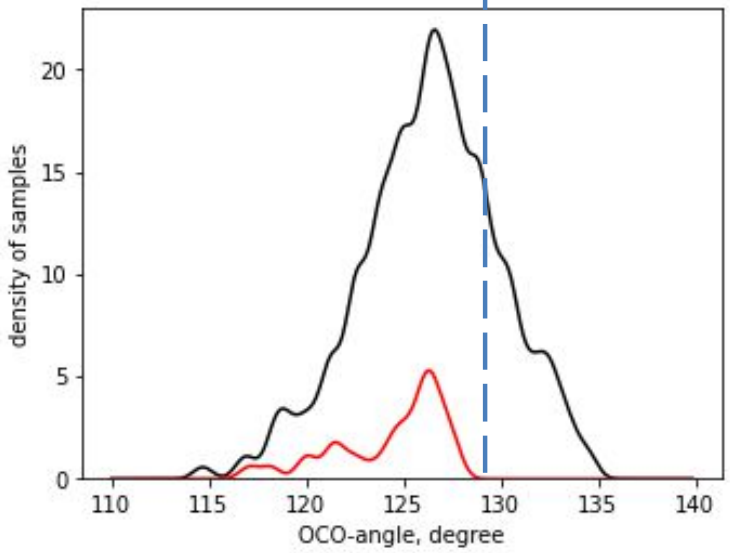
Longer C-O implies smaller OCO angles, but not too small → no catalyst poisoning

Subgroup discovery: A different approach

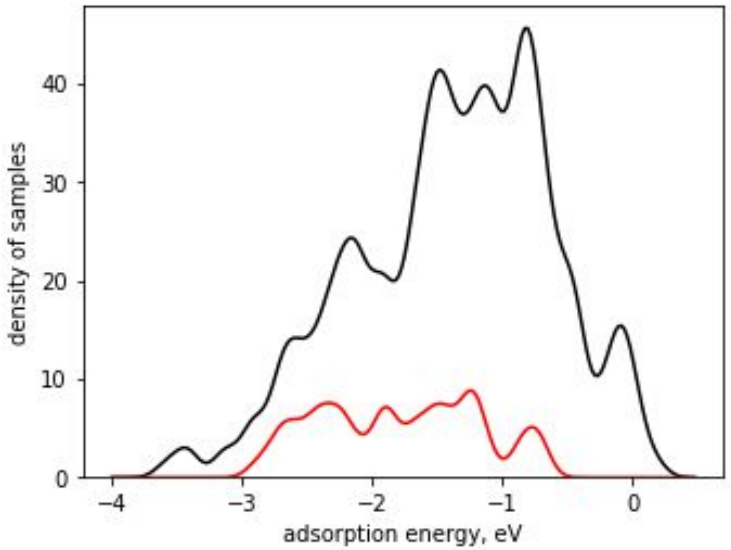
$$f = \left(\frac{N_{\text{subgroup}}}{N_{\text{all}}}\right)\left(\frac{mean_{\text{subgroup}} - mean_{\text{all}}}{min/max_{\text{all}} - mean_{\text{all}}}\right)\left(1 - \frac{variance_{\text{subgroup}}}{variance_{\text{all}}}\right)$$

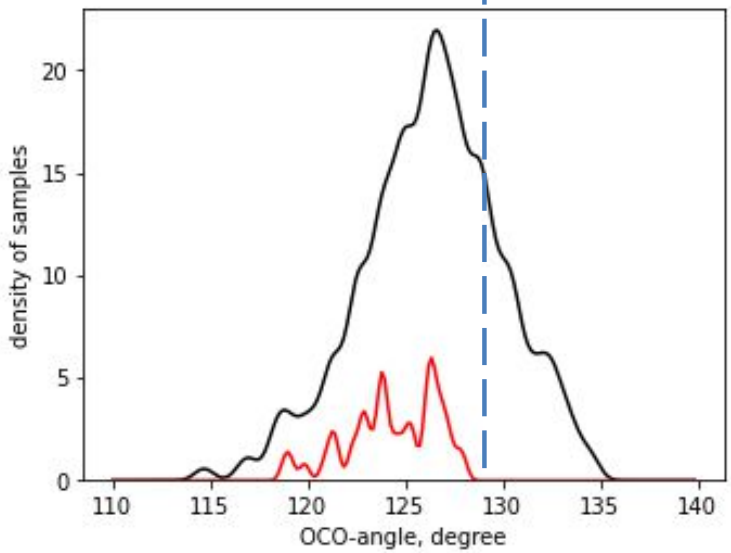
$$f = \left(\frac{N_{\text{subgroup}}}{N_{\text{all}}}\right)\left(\frac{min/max_{\text{subgroup}} - target_{\text{gas}}}{min/max_{\text{all}} - target_{\text{gas}}}\right)[\Theta(OCO < a)][\text{information gain}(E_{\text{ads}} \in (-2.5, -0.5))]$$



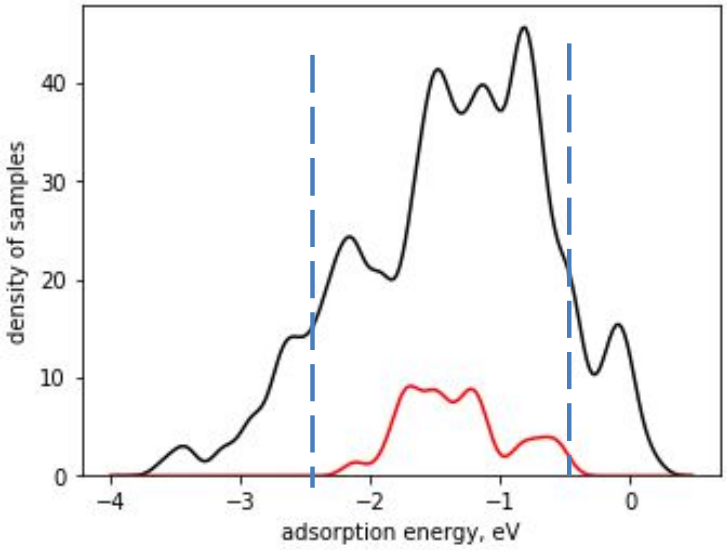


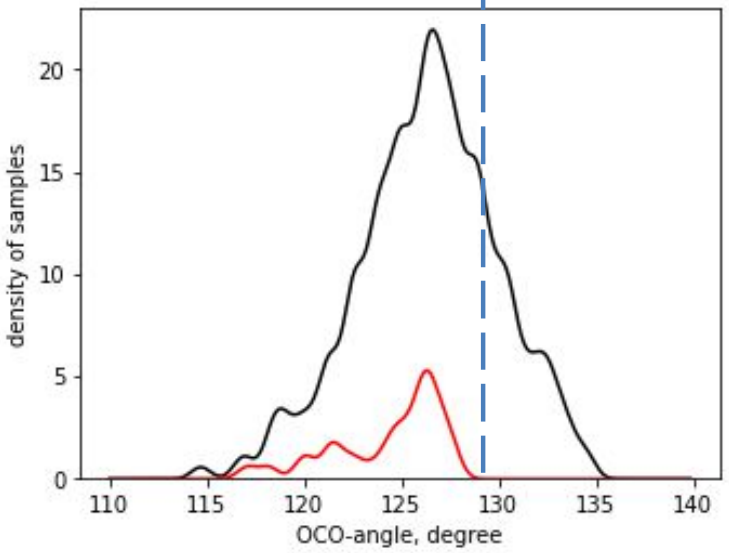
Subgroup discovery: A different approach





large Hirshfeld charge on surface O, lower coordination for smaller angles





with adsorption energy constraint:

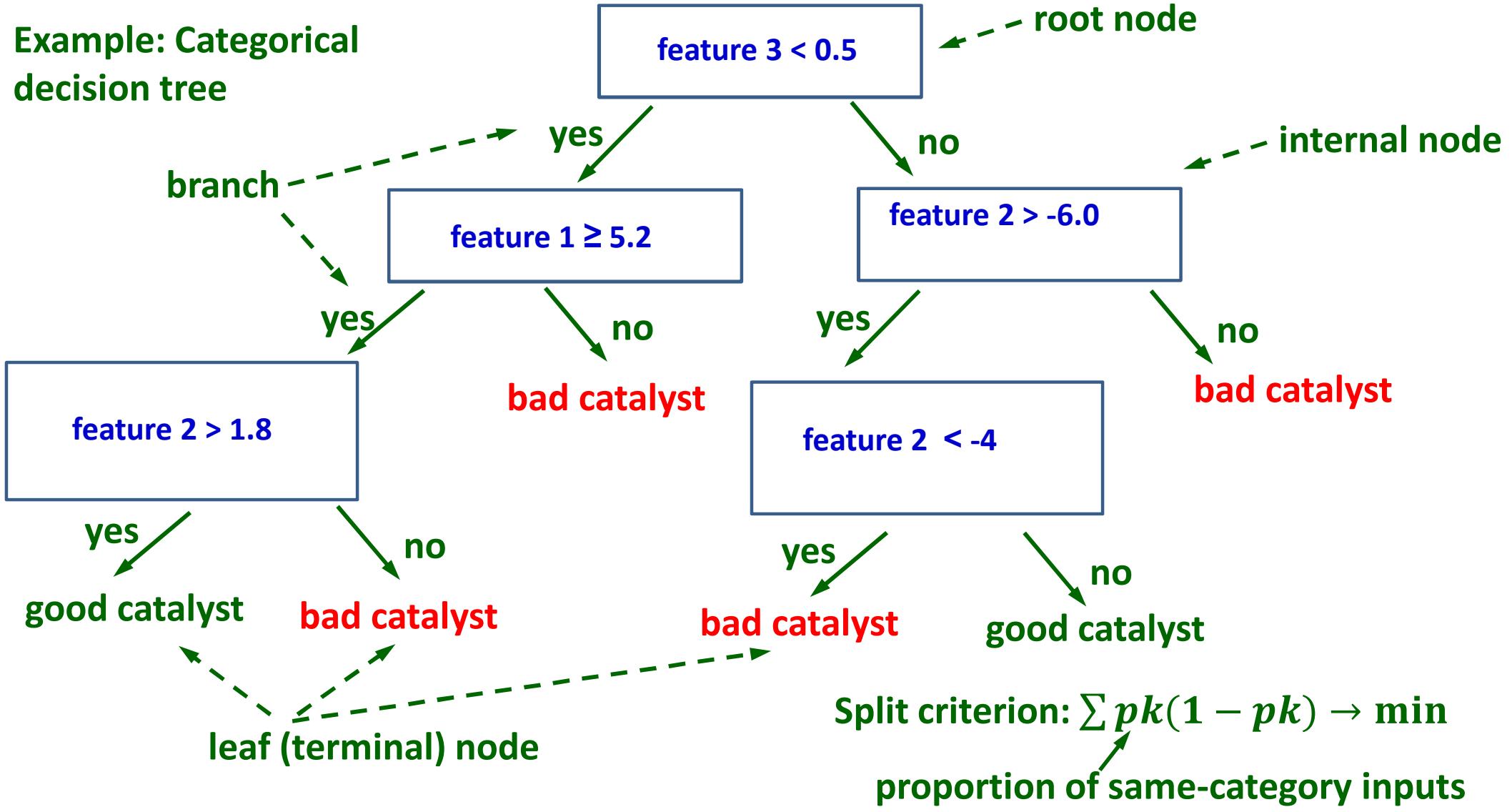
smaller charge on surface O, delocalized electron density, binding of O in CO_2 with surface cations

SISSO and SGD software

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

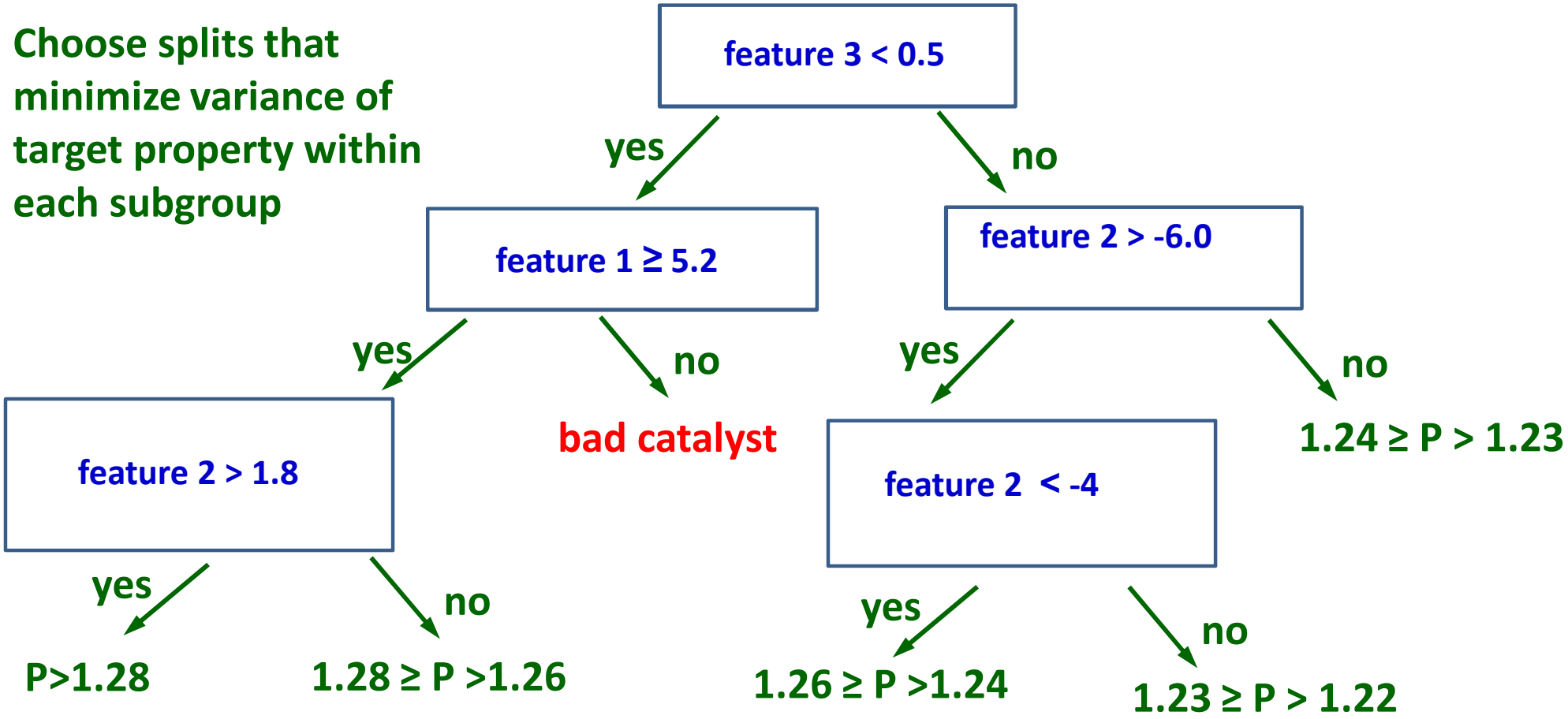
Subgroup discovery: http://www.realkd.org/

Decision trees



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 use to select primary features for other approaches such as SISSO

Random forest®

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

THE JOURNAL OF PHYSICAL CHEMISTRY C

Article

Cite This: J. Phys. Chem. C 2019, 123, 2321−2328

pubs.acs.org/JPCC

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

Robert B. Wexler,[†] Tian Qiu,[†] and Andrew M. Rappe*

JACS

JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

Cite This: J. Am. Chem. Soc. 2018, 140, 4678−4683

pubs.acs.org/JACS

Chemical Pressure-Driven Enhancement of the Hydrogen Evolving Activity of Ni_2P from Nonmetal Surface Doping Interpreted via Machine Learning

Robert B. Wexler,[†] John Mark P. Martirez,[‡] and Andrew M. Rappe*,[†]