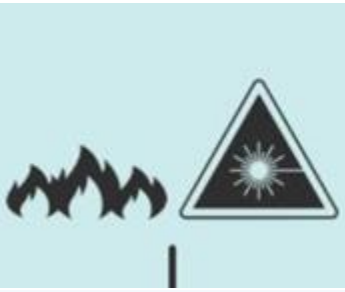
Computational Chemistry and Materials Modeling:

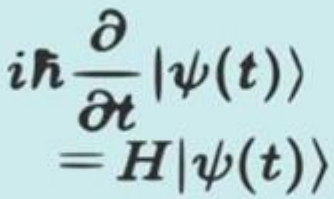
Exploring Materials Space

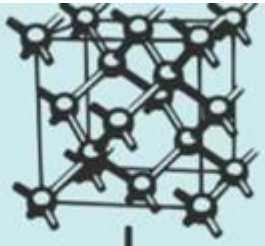
Sergey Levchenko

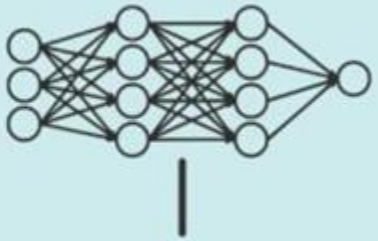
Center for Energy Science and Technology (CEST)
Skolkovo Institute of Science and Technology
Moscow, Russia

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.

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

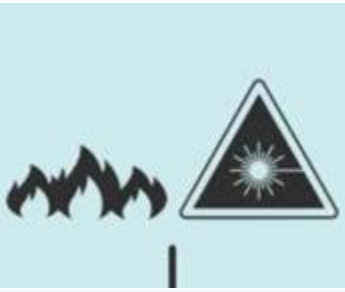
data

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

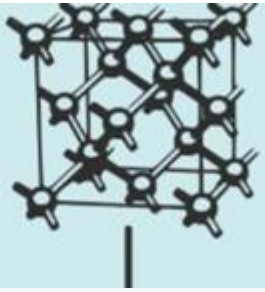
data

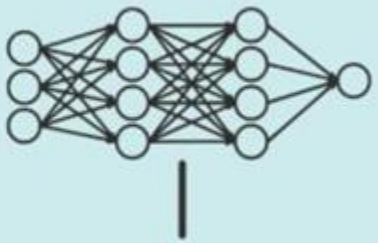
target properties (catalytic reaction yield, hardness, superconductor critical temperature, ...)

Research paradigm shift



$$i\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = H|\psi(t)\rangle$$





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:

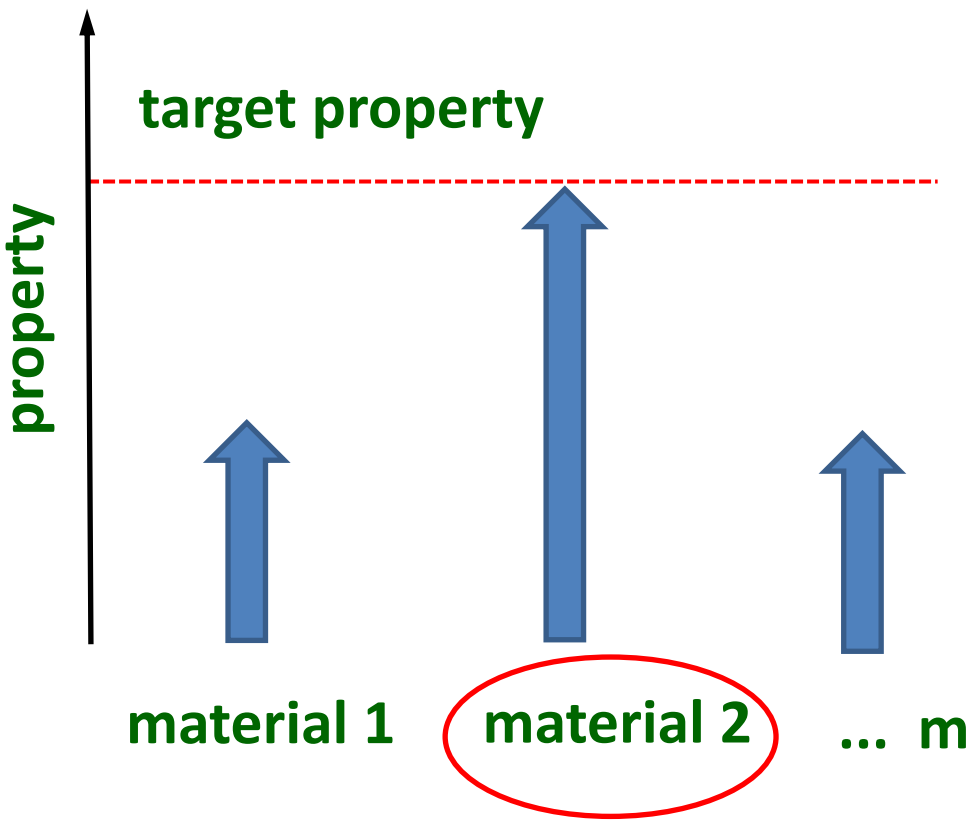
target property (high activity and selectivity of a catalyst)

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

synthesis recipe

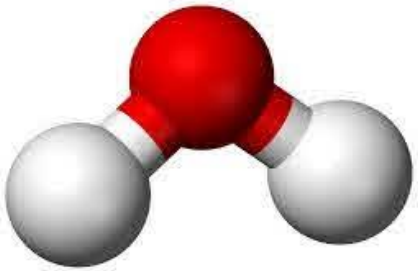
not clear how to achieve this!

Bottom-up design:

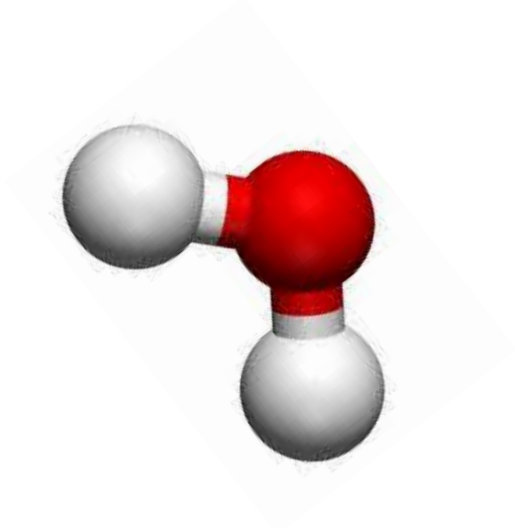


high-throughput screening

Descriptors



molecule transfer
and rotation



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

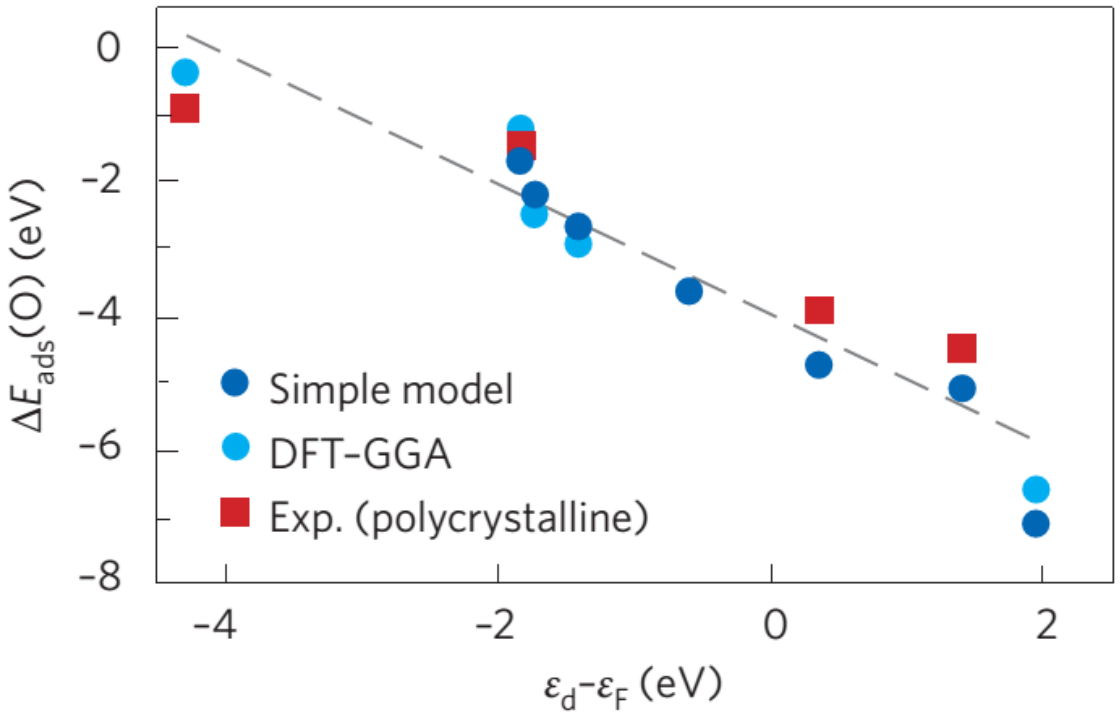
descriptive parameters (composition, synthesis conditions, operation conditions)

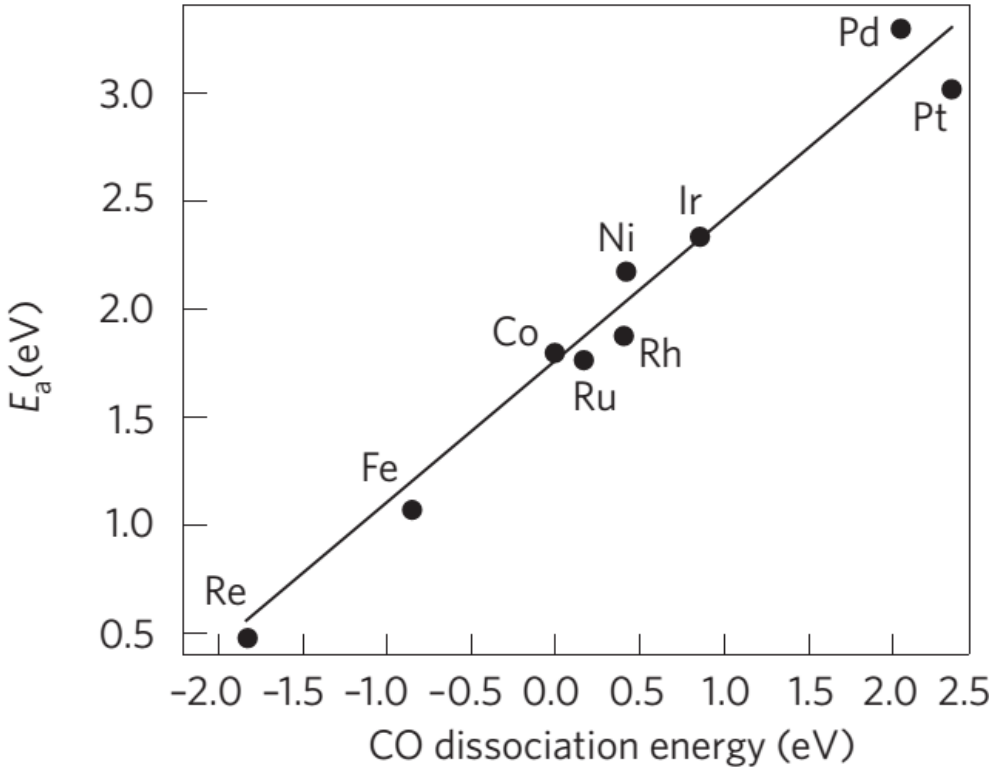
→

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

machine will learn symmetries, not (other) physics

Descriptors

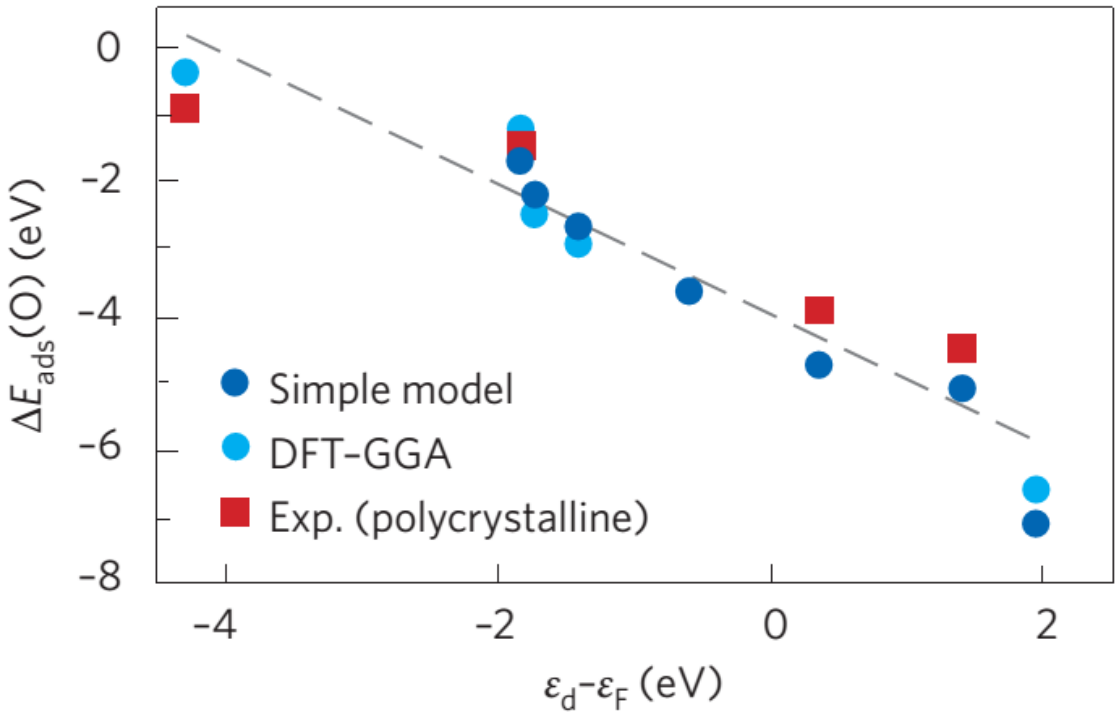


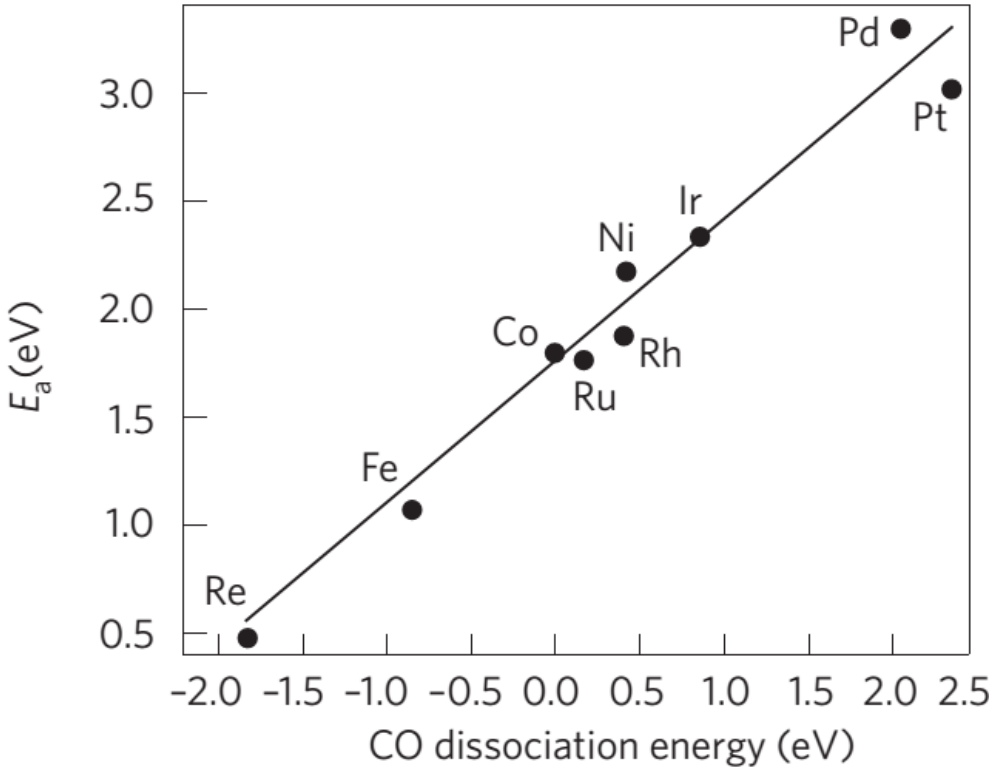


Simple(r) properties (bulk d-band center position and CO dissociation energy) are correlated to more complex properties (adsorption energy and reaction barrier)

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

Descriptors



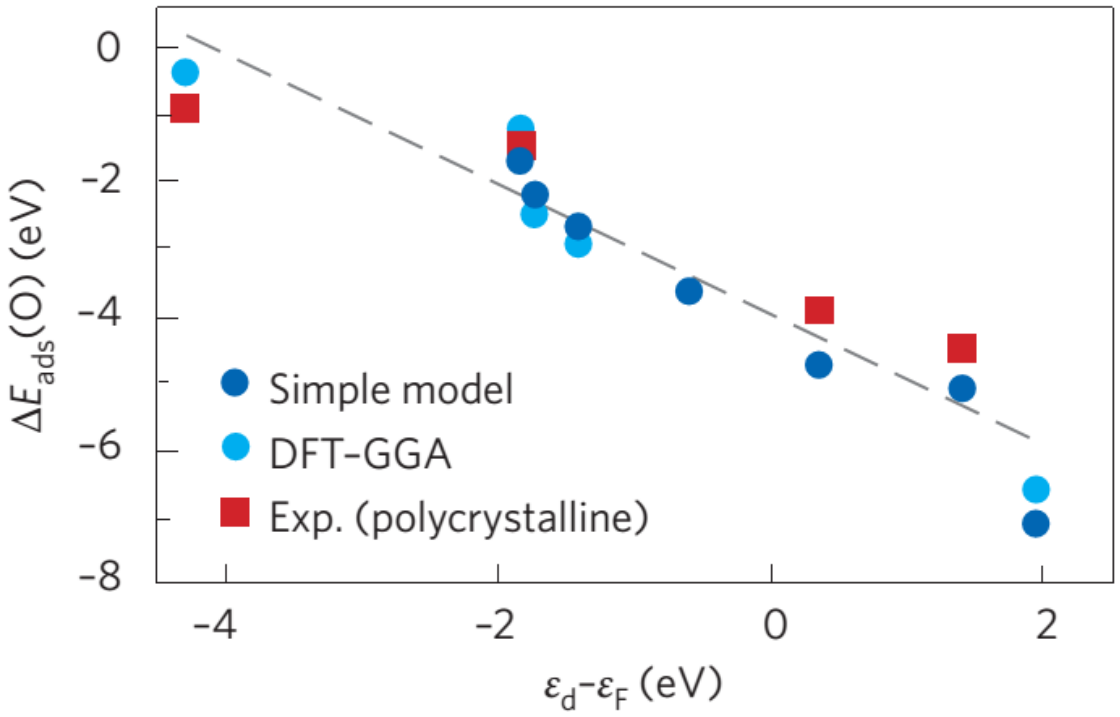


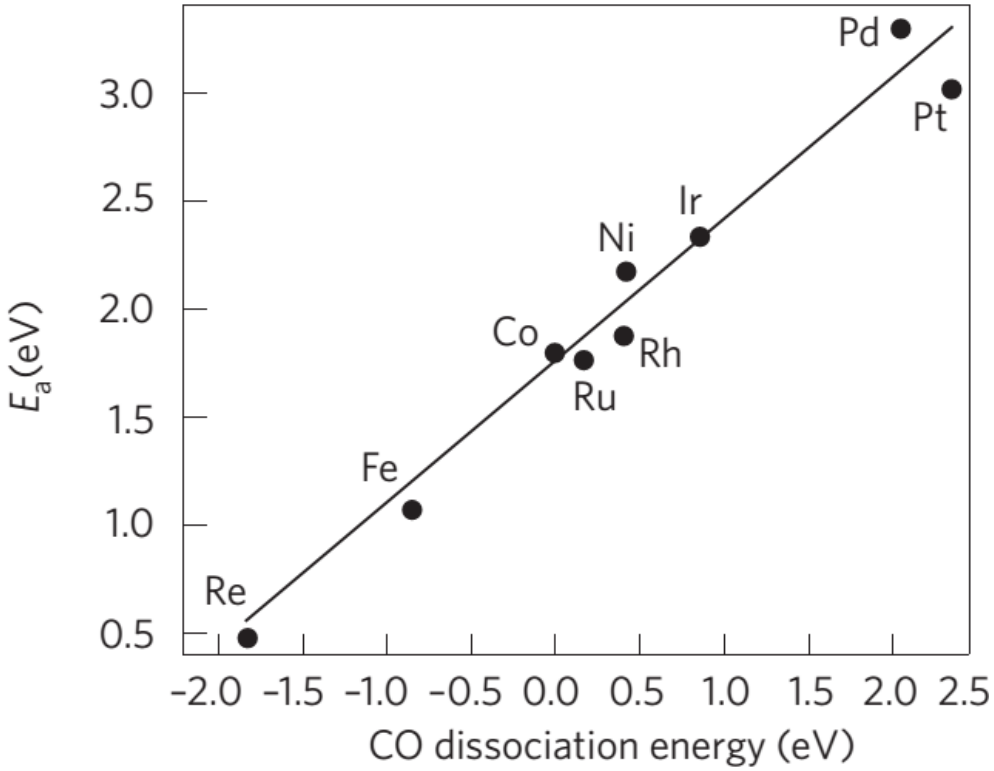
A simple physical model (Newns-Anderson) motivates the *d*-band center descriptor

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



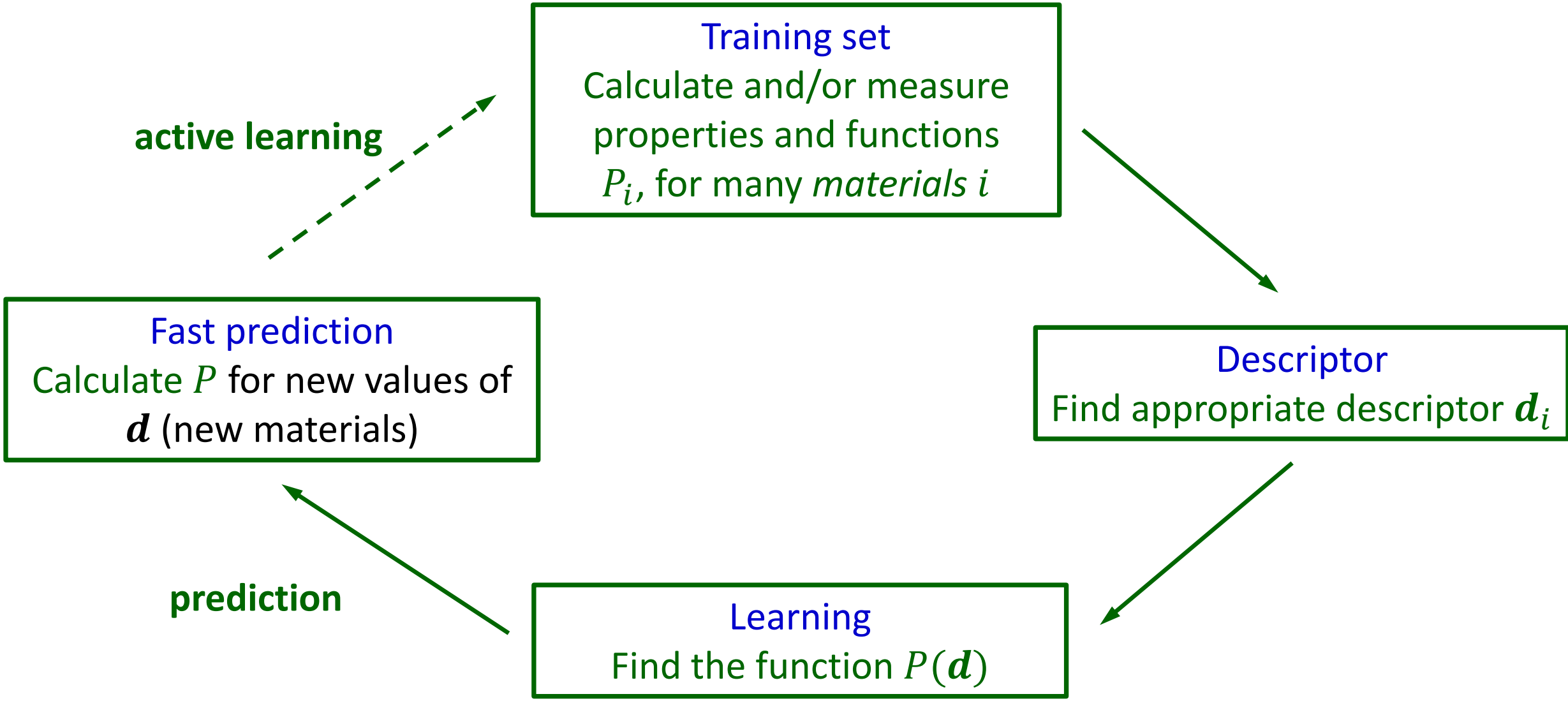


A simple physical model (Newns-Anderson) motivates the *d*-band center descriptor

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 or measurements 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
$P(\boldsymbol{d}) = \sum_{i=1}^{N} c_i \exp\left(-\lVert \boldsymbol{d}_i - \boldsymbol{d} \rVert_2^2 / 2\sigma^2\right)$	$P(\boldsymbol{d}) = \boldsymbol{d}\boldsymbol{c}$
minimize	
$\sum_{i=1}^{N} (P(\boldsymbol{d}_i) - P_i)^2 \quad +$	$\sum_{i=1}^{N} (P(\boldsymbol{d}_i) - P_i)^2 \quad +$
$\lambda \sum_{i,j=1}^{N,N} c_i c_j \exp\left(-\lVert \boldsymbol{d}_i - \boldsymbol{d}_j \rVert_2^2 / 2\sigma^2\right)$	$\lambda \lVert \boldsymbol{c} \rVert_0$
$\lVert \boldsymbol{d}_i - \boldsymbol{d}_j \rVert_2^2 = \sum_{\alpha=1}^{\Omega} (d_{i,\alpha} - d_{j,\alpha})^2$	

Target property model: Kernel ridge regression versus feature selection

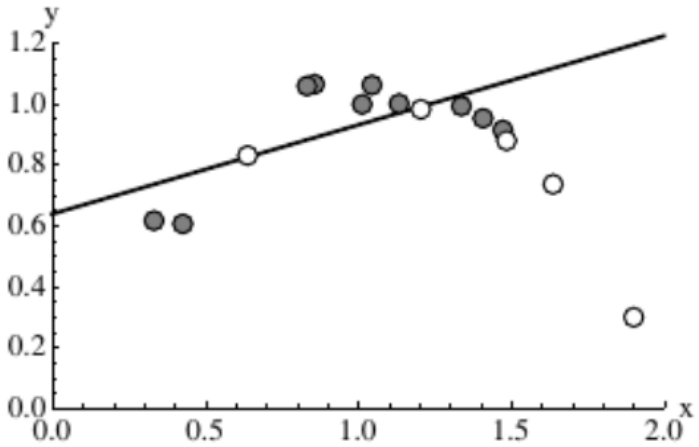
kernel (Gaussian, Laplacian, linear $(\boldsymbol{d}_i \cdot \boldsymbol{d}_j)$)

kernel ridge regression		linear
$P(\boldsymbol{d}) = \sum_{i=1}^{N} c_i \exp\left(-\|\boldsymbol{d}_i - \boldsymbol{d}\|_2^2/2\sigma^2\right)$		$P(\boldsymbol{d}) = \boldsymbol{d}\boldsymbol{c}$
$\sum_{i=1}^{N}(P(\boldsymbol{d}_i) - P_i)^2 \quad +$	minimize	$\sum_{i=1}^{N}(P(\boldsymbol{d}_i) - P_i)^2 \quad +$
$\underbrace{\lambda \sum_{i,j=1}^{N,N} c_i c_j \exp\left(-\|\boldsymbol{d}_i - \boldsymbol{d}_j\|_2^2/2\sigma^2\right)}_{\text{penalty on similar data points}}$		$\underbrace{\lambda\|\boldsymbol{c}\|_0}_{\text{penalty on the number of non-zero coefficients } \|\boldsymbol{c}\|_0}$

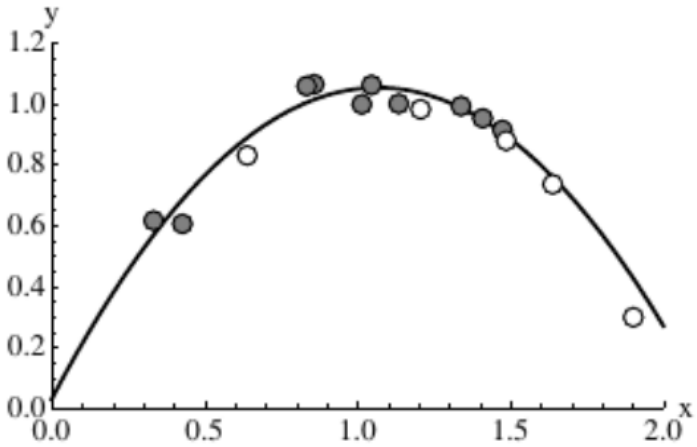
Regression: Importance of regularization

● training ○ validation

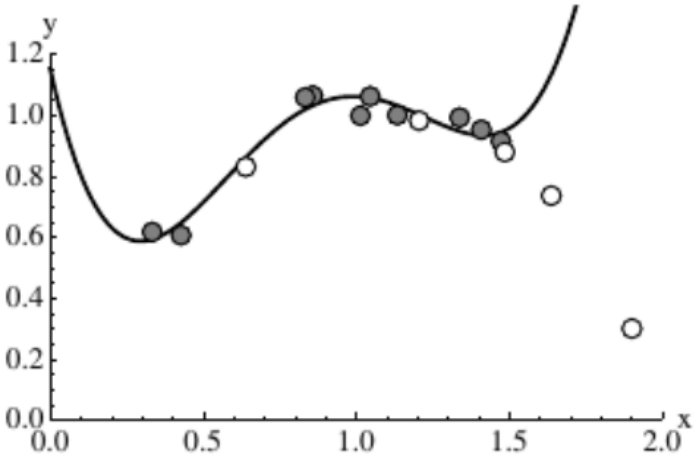
Underfitting



Fitting



Overfitting



Training/ validation error

0.123 / 0.443

0.044 / 0.068

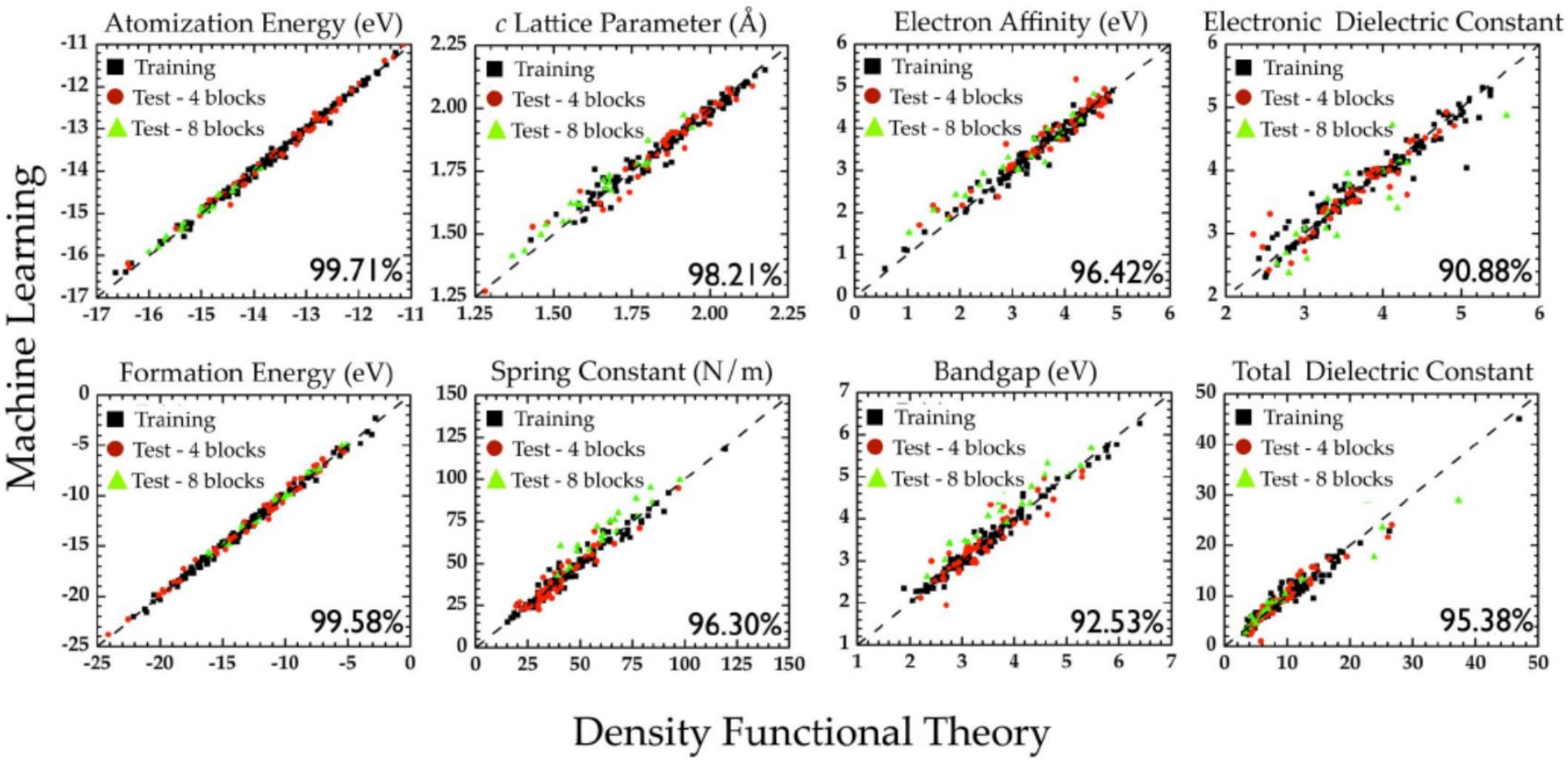
0.036 / 0.939

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

(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**

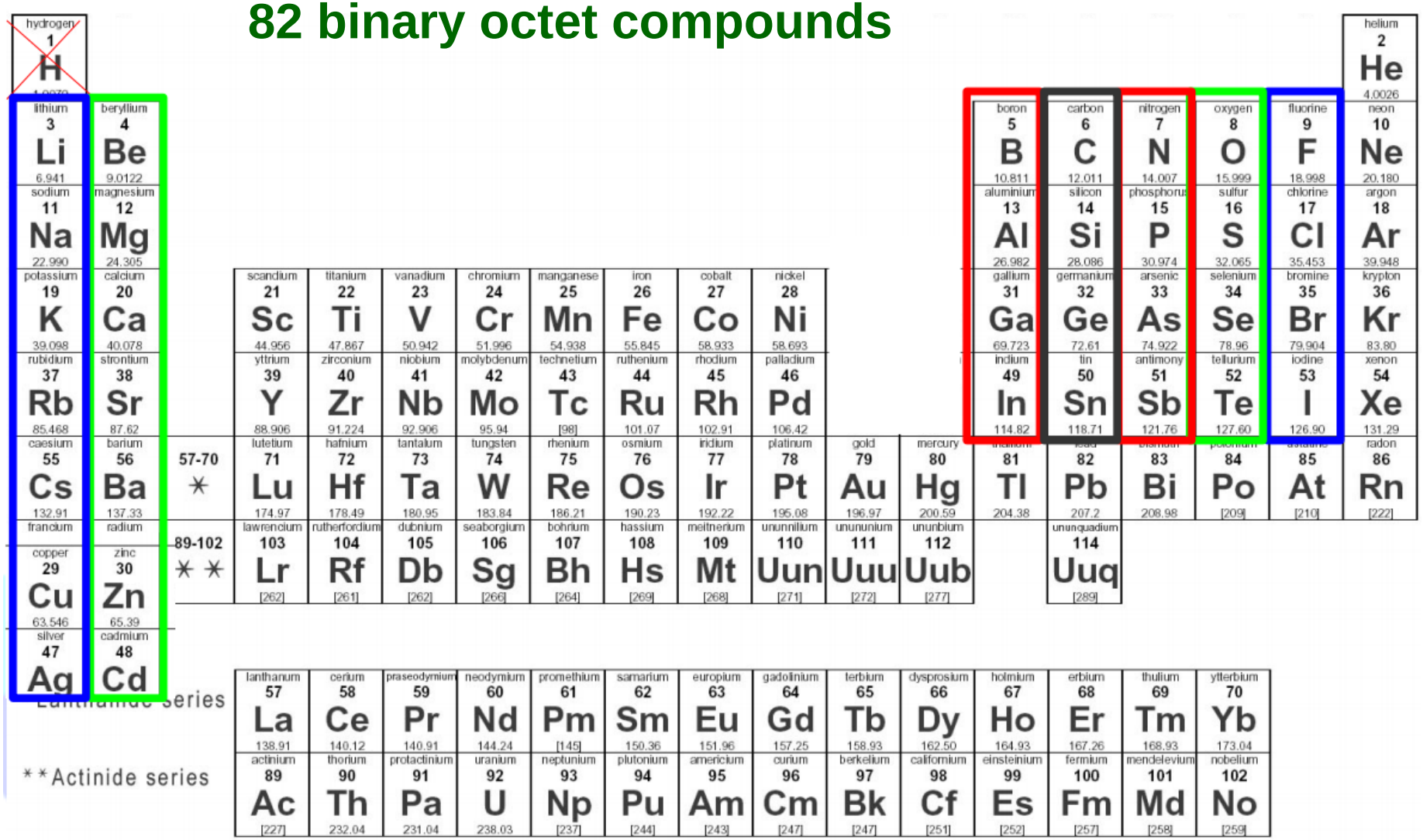
3) **The dimension Ω of the descriptor should be as low as possible (for a certain accuracy request)**

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

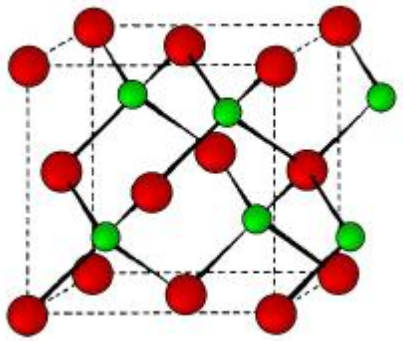
L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Phys. Rev. Lett. **114**, 105503 (2015)

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

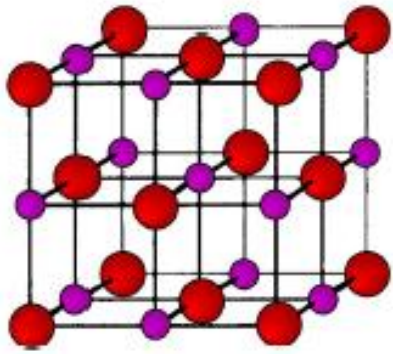
82 binary octet compounds



ZB



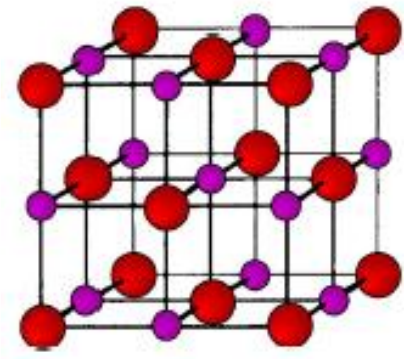
RS



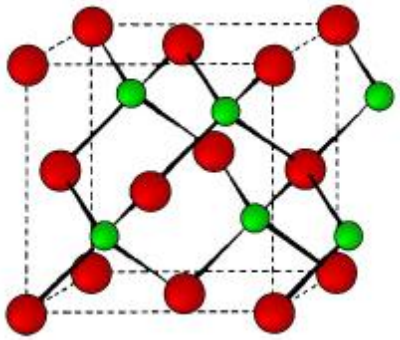
Proof of Concept: Descriptor for the Classification "Zincblende/Wurtzite (ZB/W) or Rocksalt (RS)?"

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

RS



ZB



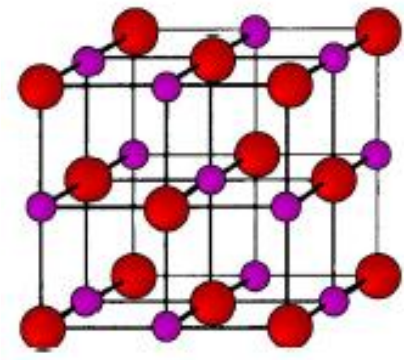
Energy differences between different structures are very small.

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

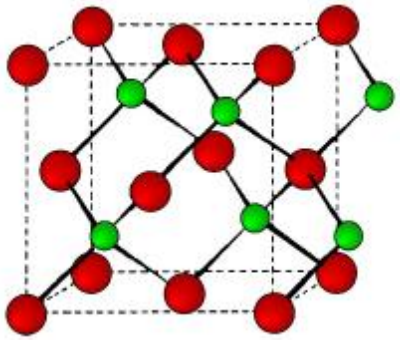
Proof of Concept: Descriptor for the Classification "Zincblende/Wurtzite (ZB/W) or Rocksalt (RS)?"

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

RS



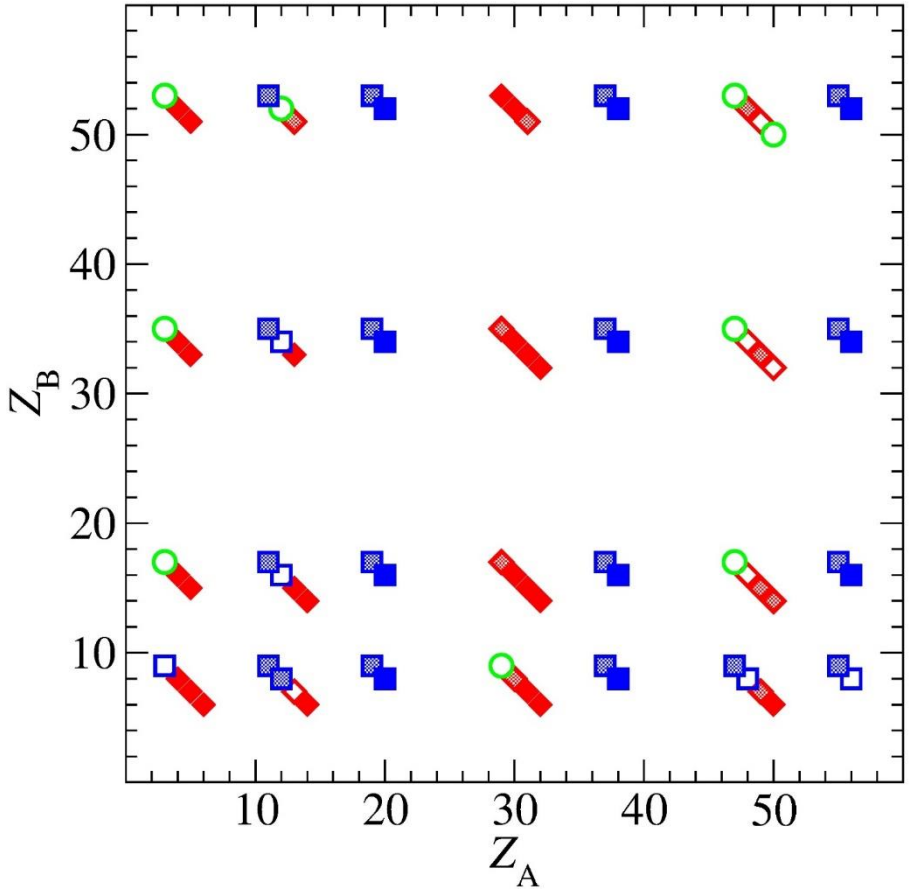
ZB

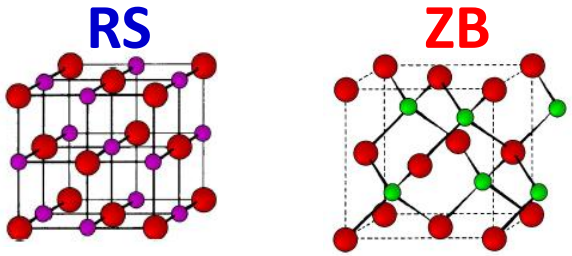


J. A. van Vechten, Phys. Rev. 182, 891 (1969). **J. C. Phillips**, Rev. Mod. Phys. 42, 317 (1970).
J. John and A.N. Bloch, Phys. Rev. Let. 33, 1095 (1974) **J. R. Chelikowsky and J. C. Phillips**, Phys. Rev. B 33, 2453 (1978)
A. Zunger, Phys. Rev. B 22, 5839 (1980).
D. G. Petifor, Solid State Commun.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





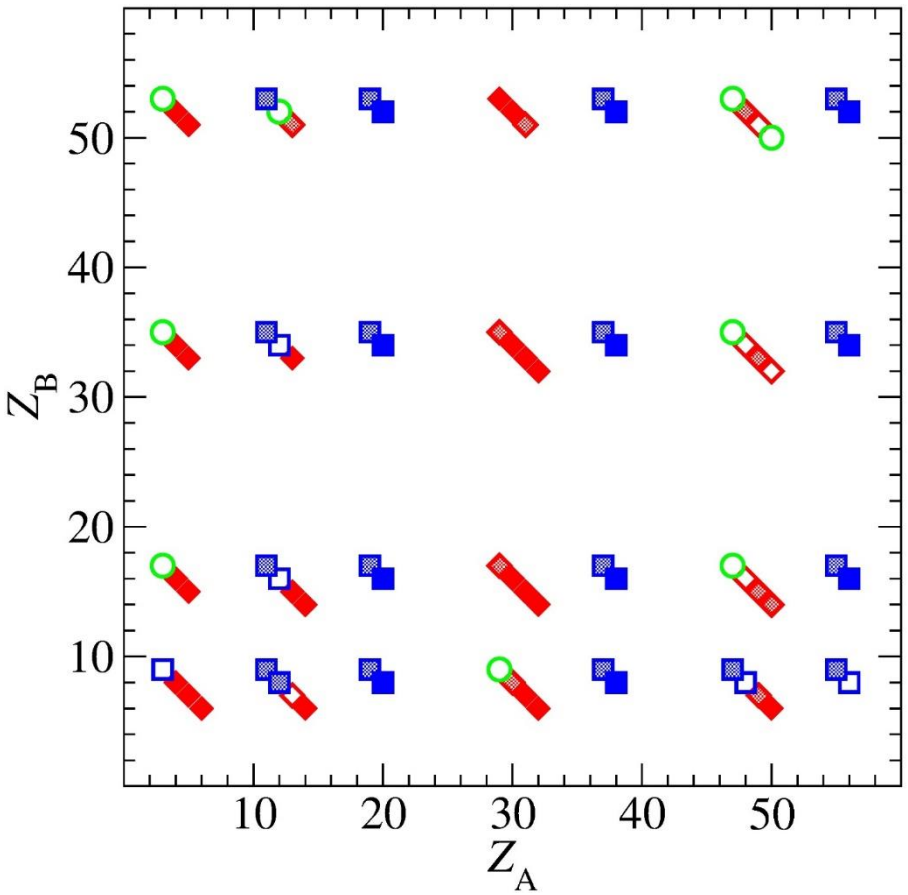
$\Delta = E(\mathrm{RS}) - E(\mathrm{ZB})$

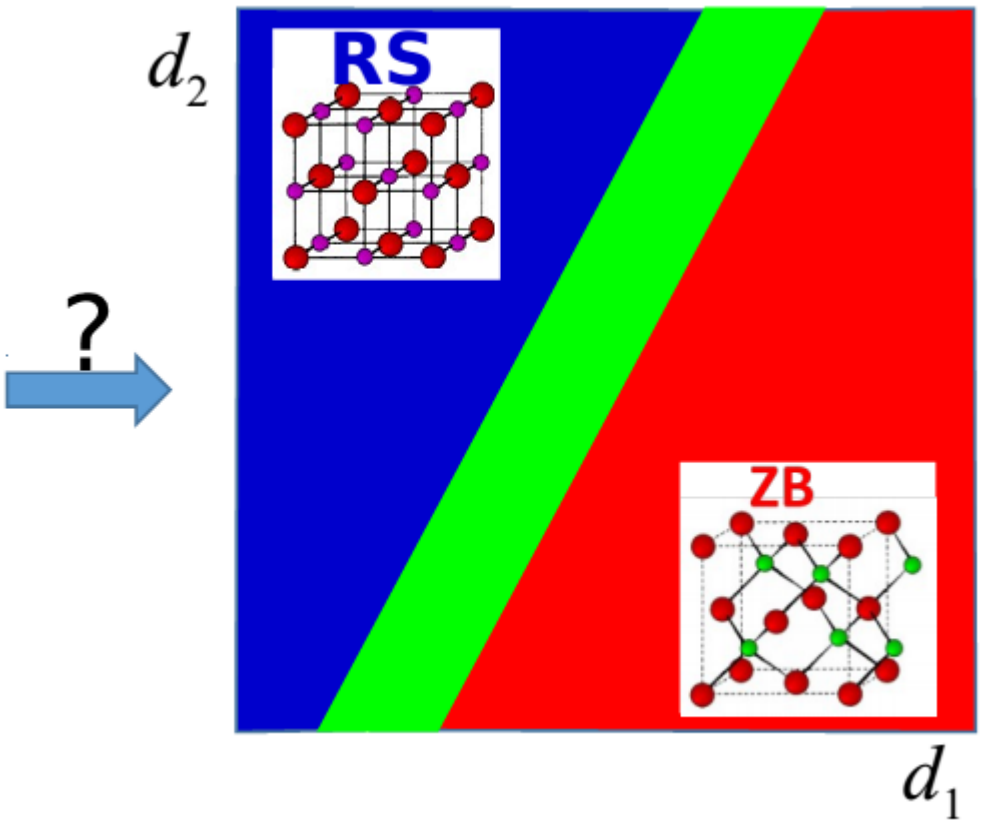
- ZB, $\Delta > 0.2$ eV
- ZB, $0.1\ \mathrm{eV} < \Delta \leq 0.2$ eV
- ZB, $0.05\ \mathrm{eV} < \Delta \leq 0.1$ eV
- $-0.05\ \mathrm{eV} < \Delta \leq 0.05$ eV
- RS, $-0.1\ \mathrm{eV} < \Delta \leq -0.05$ eV
- RS, $-0.2\ \mathrm{eV} < \Delta \leq -0.1$ eV
- RS, $\Delta \leq -0.2$ eV

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?"**

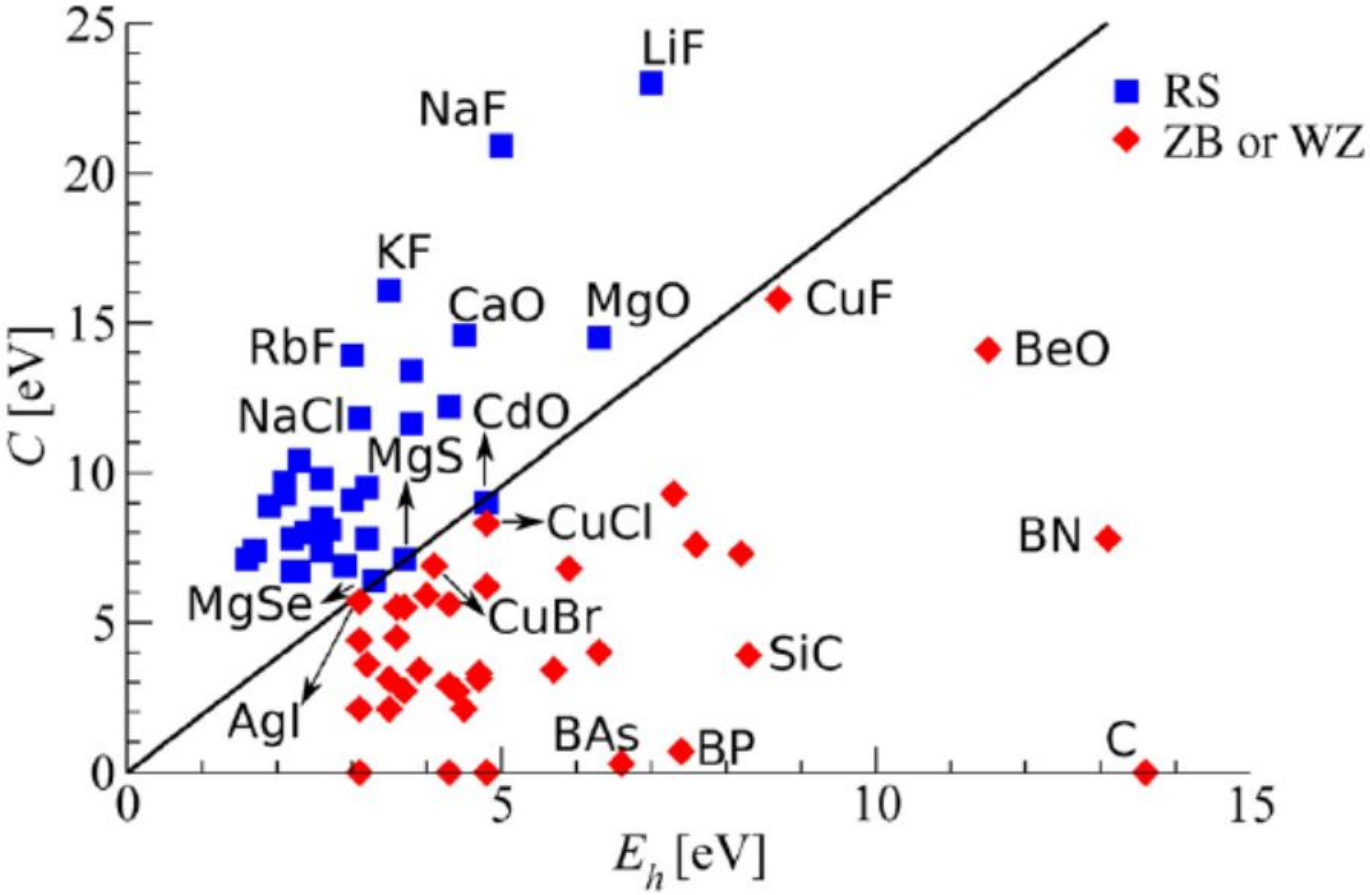
No complexity reduction → need a better basis





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?"**



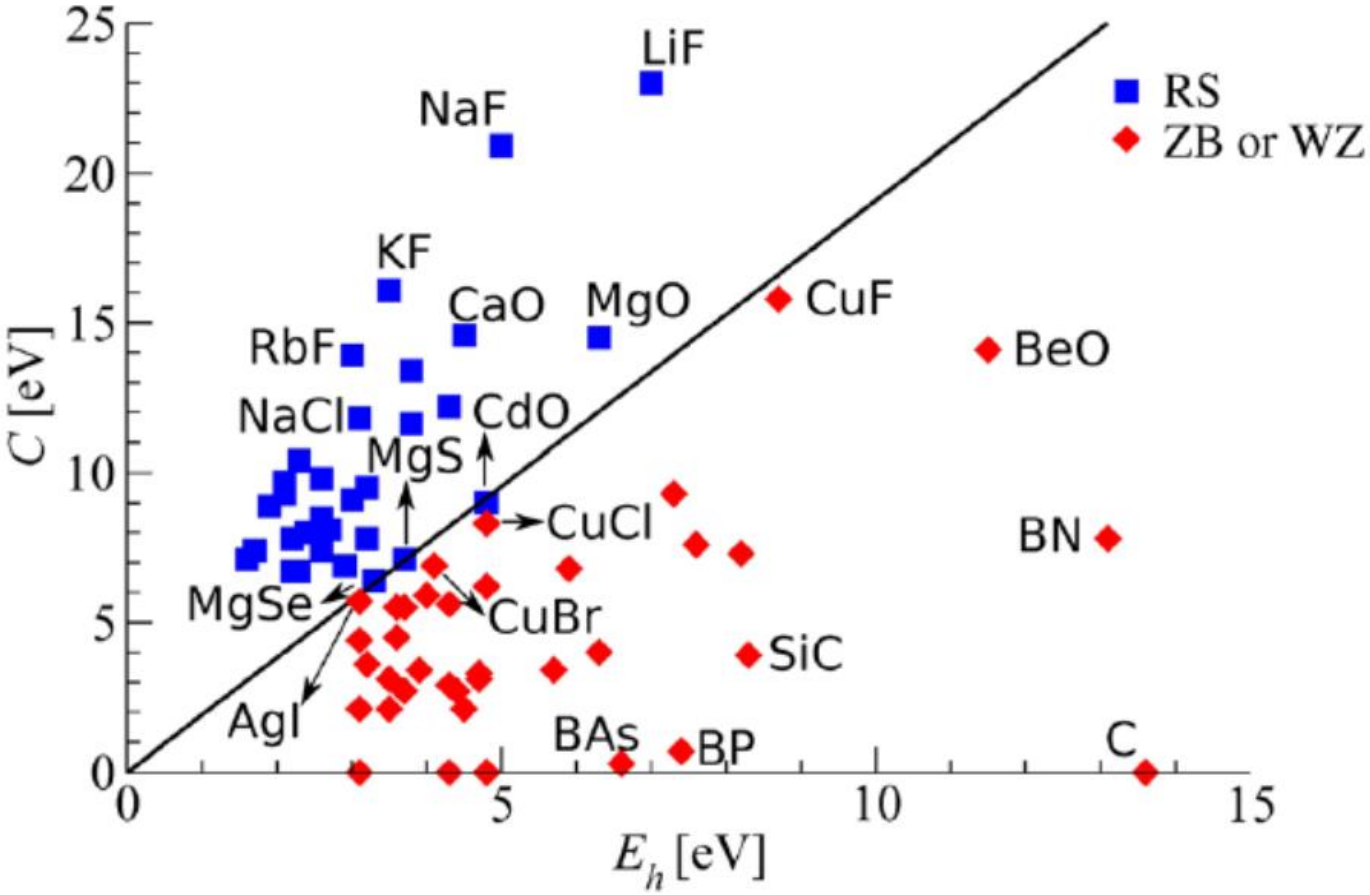
descriptor can be determined spectroscopically - properties of the solid

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?

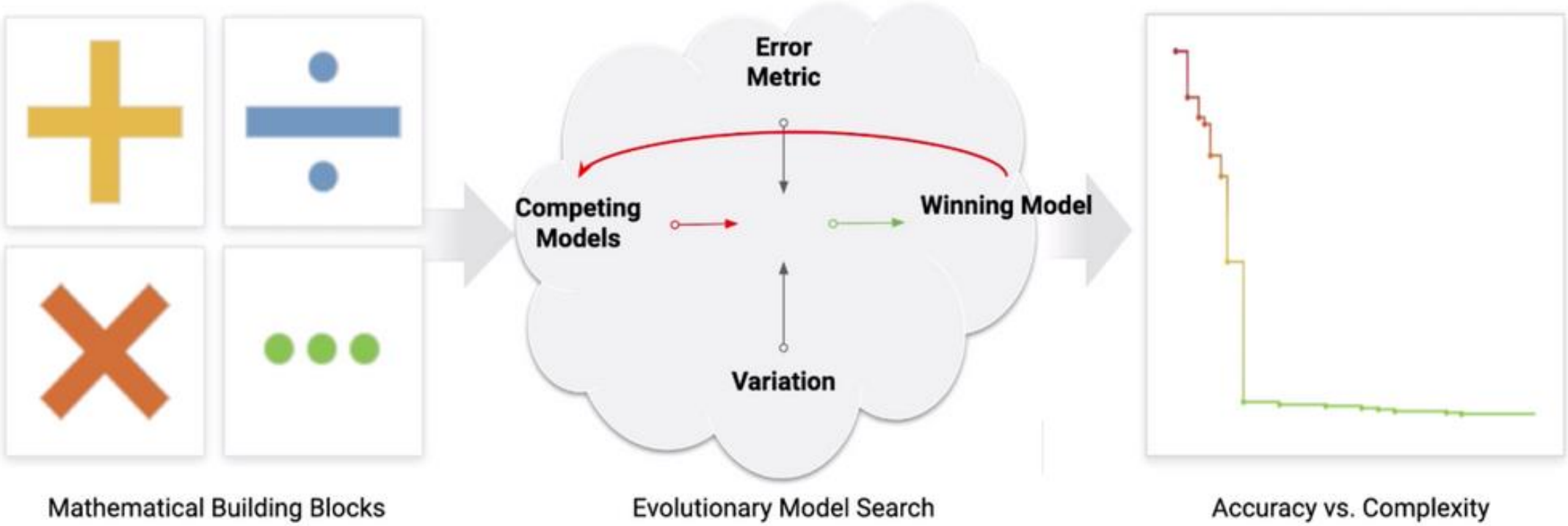
Primary features and feature space

ID	Description	Symbols	#
	free atoms		
$A1$	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

ID	Description	Symbols	#
	free dimers		
$A4$	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

ID	Description	Symbols	#
	free atoms		
$A1$	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

ID	Description	Symbols	#
	free dimers		
$A4$	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 $A1$	$\lvert \mathrm{IP(A)} \pm \mathrm{IP(B)} \rvert$	12
$B2$	absolute differences and sums of $A2$	$\lvert \mathrm{L(B)} \pm \mathrm{H(A)} \rvert$	12
$B3$	absolute differences and sums of $A3$	$\lvert r_p(\mathrm{A}) \pm r_s(\mathrm{A}) \rvert$	30
$C3$	squares of $A3$ and $B3$ (only sums)	$r_s(\mathrm{A})^2, (r_p(\mathrm{A}) + r_s(\mathrm{A}))^2$	21
$D3$	exponentials of $A3$ and $B3$ (only sums)	$\exp(r_s(\mathrm{A})), \exp(r_p(\mathrm{A}) \pm r_s(\mathrm{A}))$	21
$E3$	exponentials of squared $A3$ and $B3$ (only sums)	$\exp(r_s(\mathrm{A})^2), \exp(r_p(\mathrm{A}) \pm r_s(\mathrm{A})^2)$	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?

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

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?

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

$\|c\|_0$ -- 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., $\left|r_s(A_j) - r_p(B_j)\right|$) (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?

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

$\|c\|_0$ -- 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)

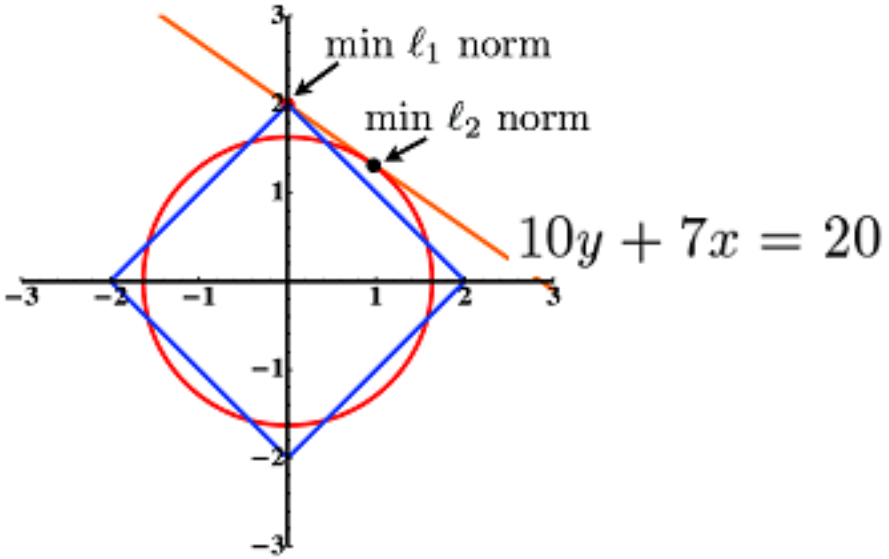
$d_{j,l}$ -- value of feature l related to material j (e.g., $|r_s(A_j) - r_p(B_j)|$) (a basis function in materials space)

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?

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



$\|\boldsymbol{c}\|_1 = \sum_l |c_l|$ -- 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

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

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

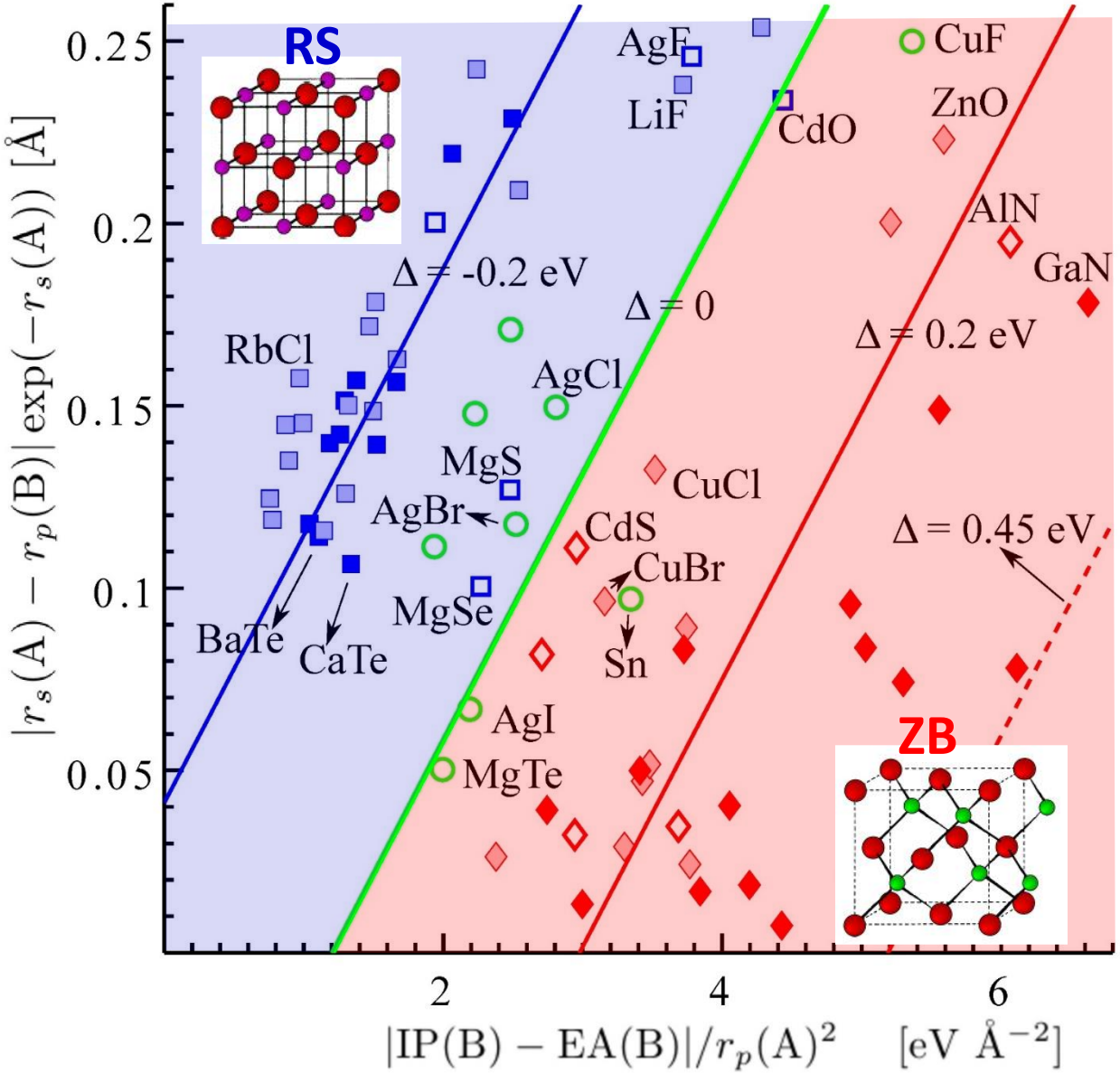
$$\Delta E = 0.113\frac{\mathrm{EA(B)} - \mathrm{IP(B)}}{r_p(\mathrm{A})^2} + 1.542\frac{|r_s(\mathrm{A}) - r_p(\mathrm{B})|}{\exp(r_s(\mathrm{A}))} - 0.137 \qquad \text{2D}$$

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

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

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

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



$\Delta = E(RS) - E(ZB)$

- ◆ ZB, $\Delta > 0.2$ eV
- ◆ ZB, 0.1 eV $< \Delta \leq 0.2$ eV
- ◇ ZB, 0.05 eV $< \Delta \leq 0.1$ eV
- ○ -0.05 eV $< \Delta \leq 0.05$ eV
- □ RS, -0.1 eV $< \Delta \leq -0.05$ eV
- ▩ RS, -0.2 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).

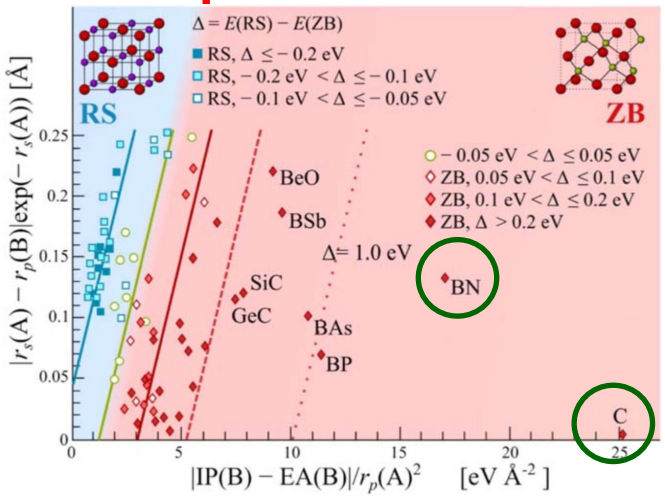
L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, and M. Scheffler, Phys. Rev. Lett. **114**, 105503 (2015).

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

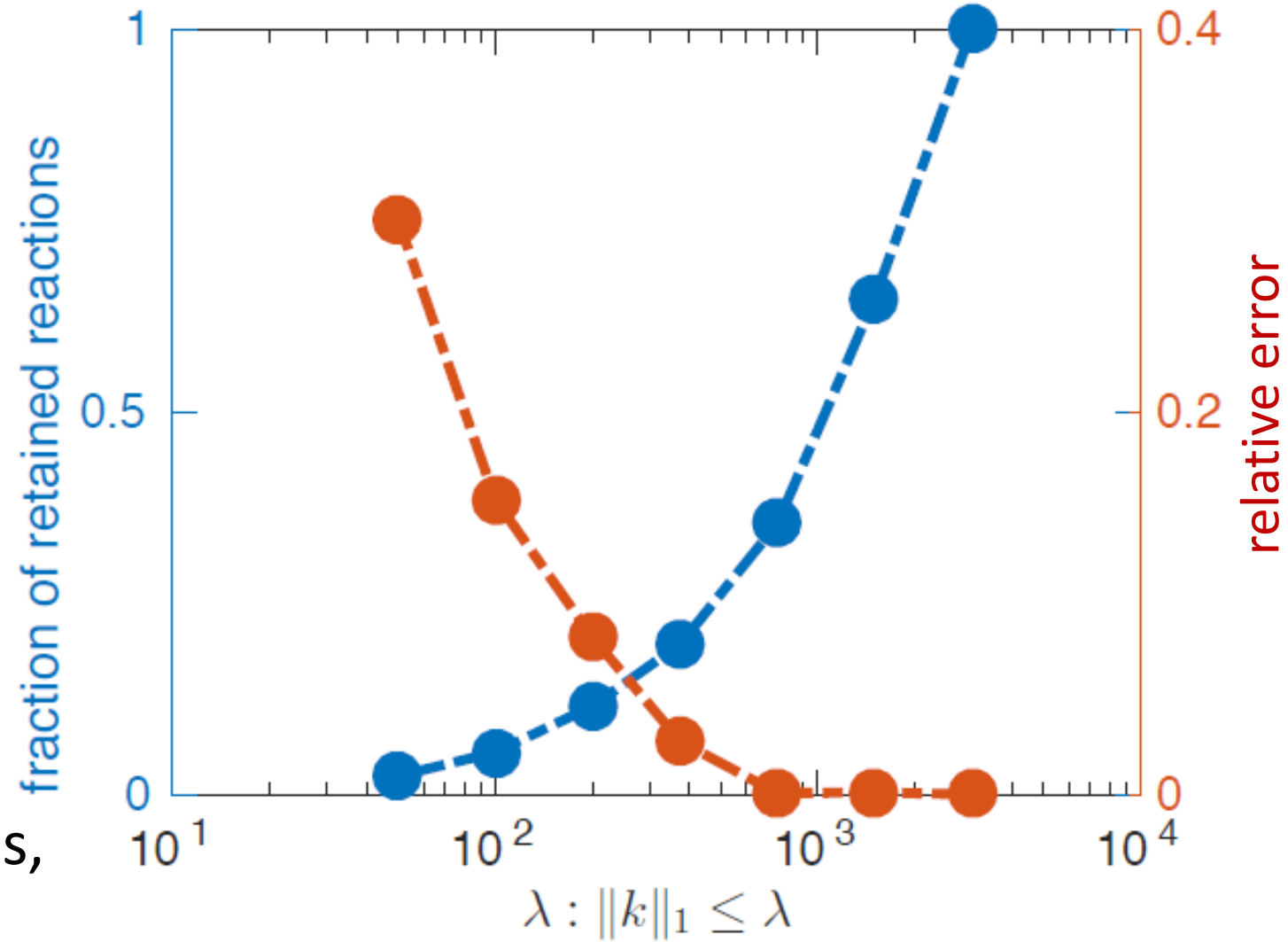
CH_4 chemical decomposition under shock-compression conditions (high *T* and *p*)

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.

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

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)

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

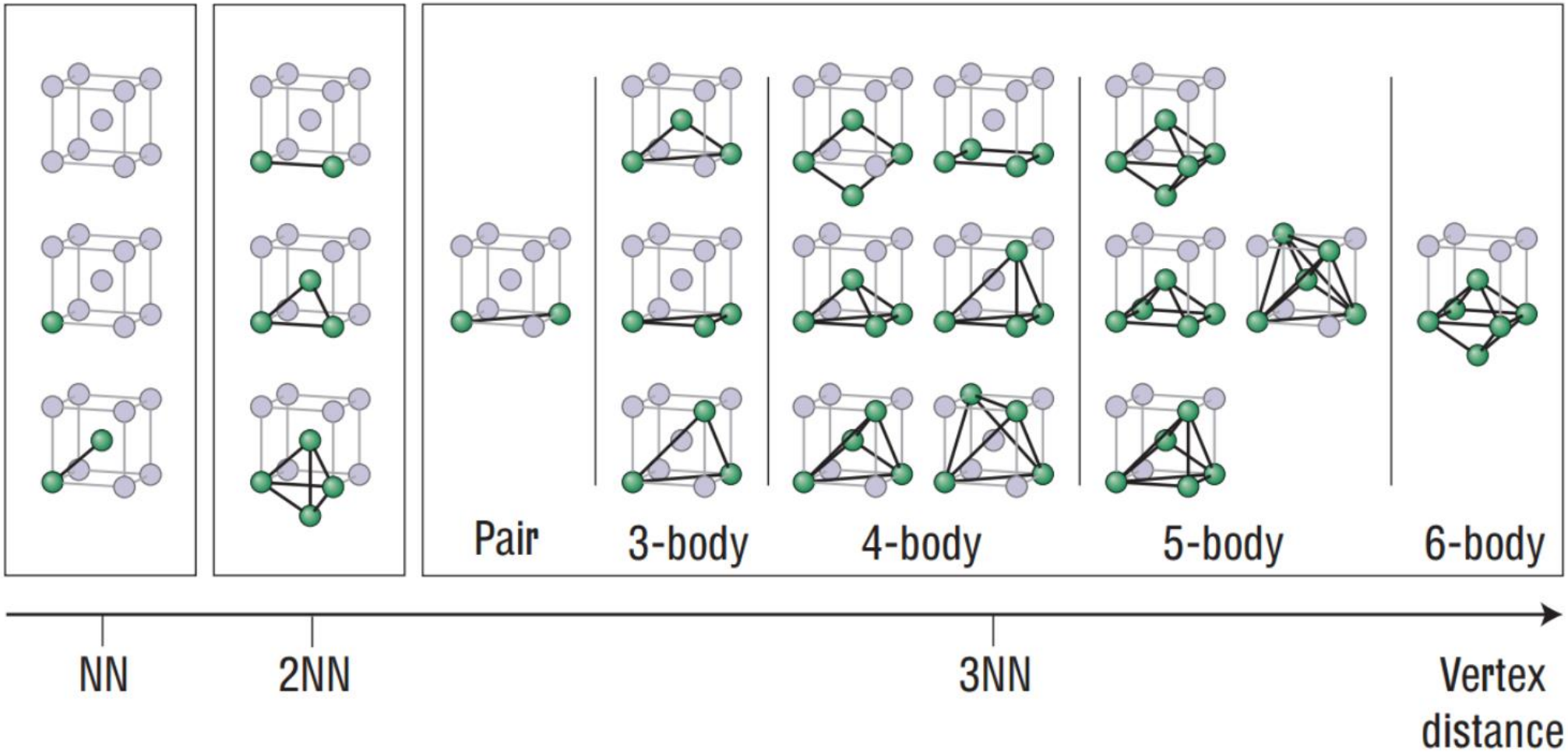
$$A_{aJ} = \begin{bmatrix} -1 & u_b^1 & -\frac{1}{2}u_b^1 u_c^1 & \cdots \\ \vdots & \vdots & \vdots & \\ -1 & u_b^L & -\frac{1}{2}u_b^L u_c^L & \cdots \end{bmatrix}$$

→ **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

$\|\boldsymbol{c}\|_1 = \sum_l |c_l|$ **-- 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 ($\boldsymbol{D}^T \boldsymbol{y}$)

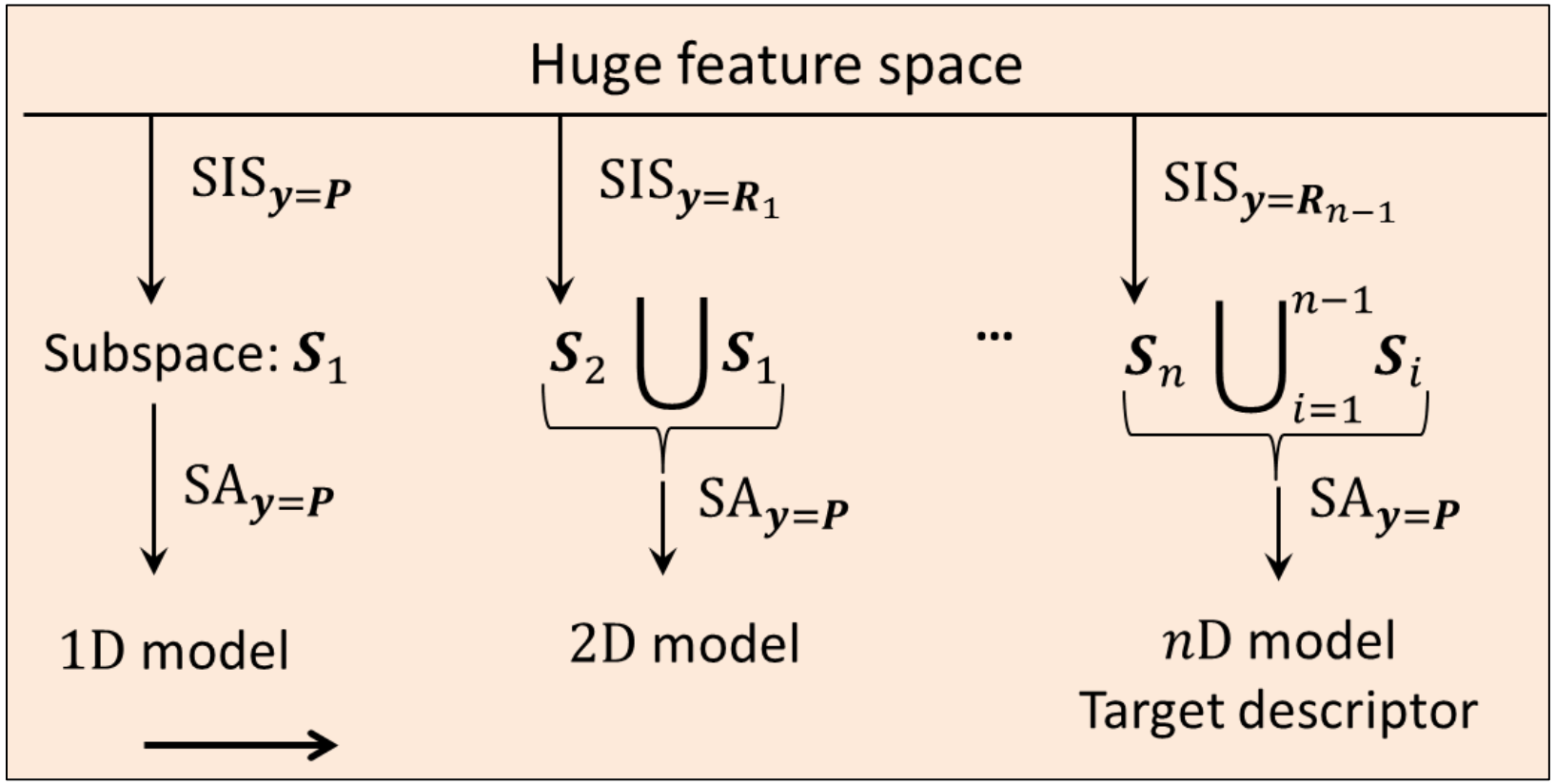
   $\boldsymbol{y}$: vector with the target property (e.g., rock salt-zincblende energy differences; 82 elements)

   $\boldsymbol{D}$: 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)

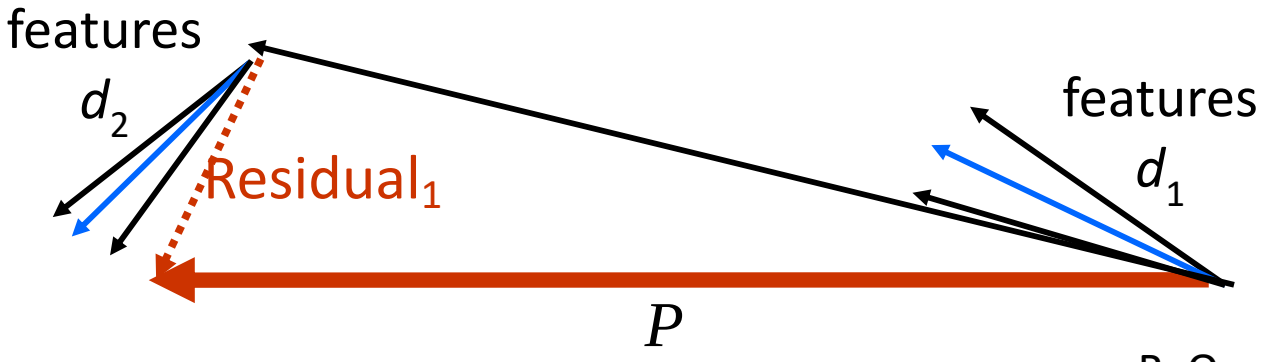
SISSO: Iterative residual fitting



$\boldsymbol{y}$: response vector

$\boldsymbol{P}$: target material property

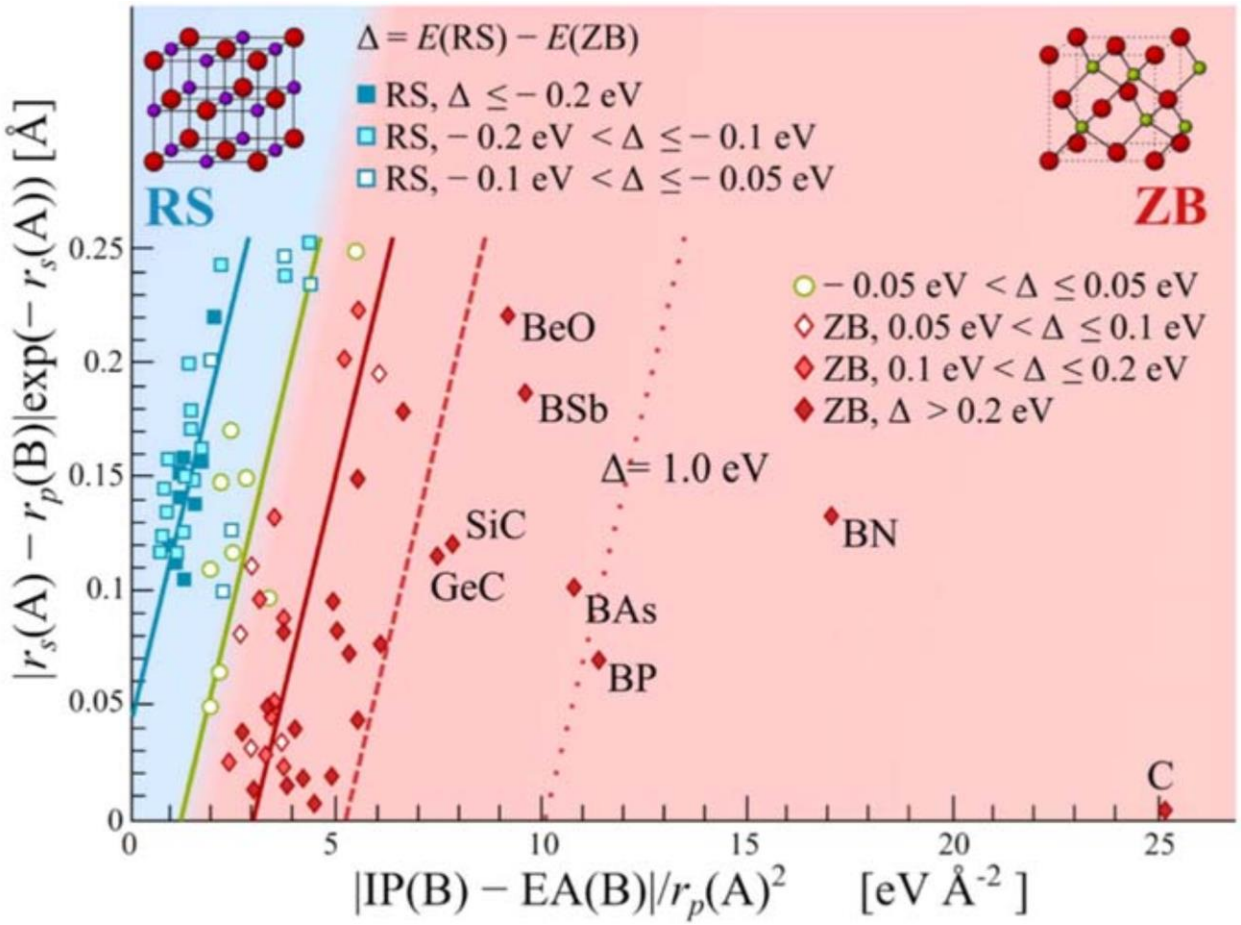
Residual: $R = P - \sum_i c_i d_i$



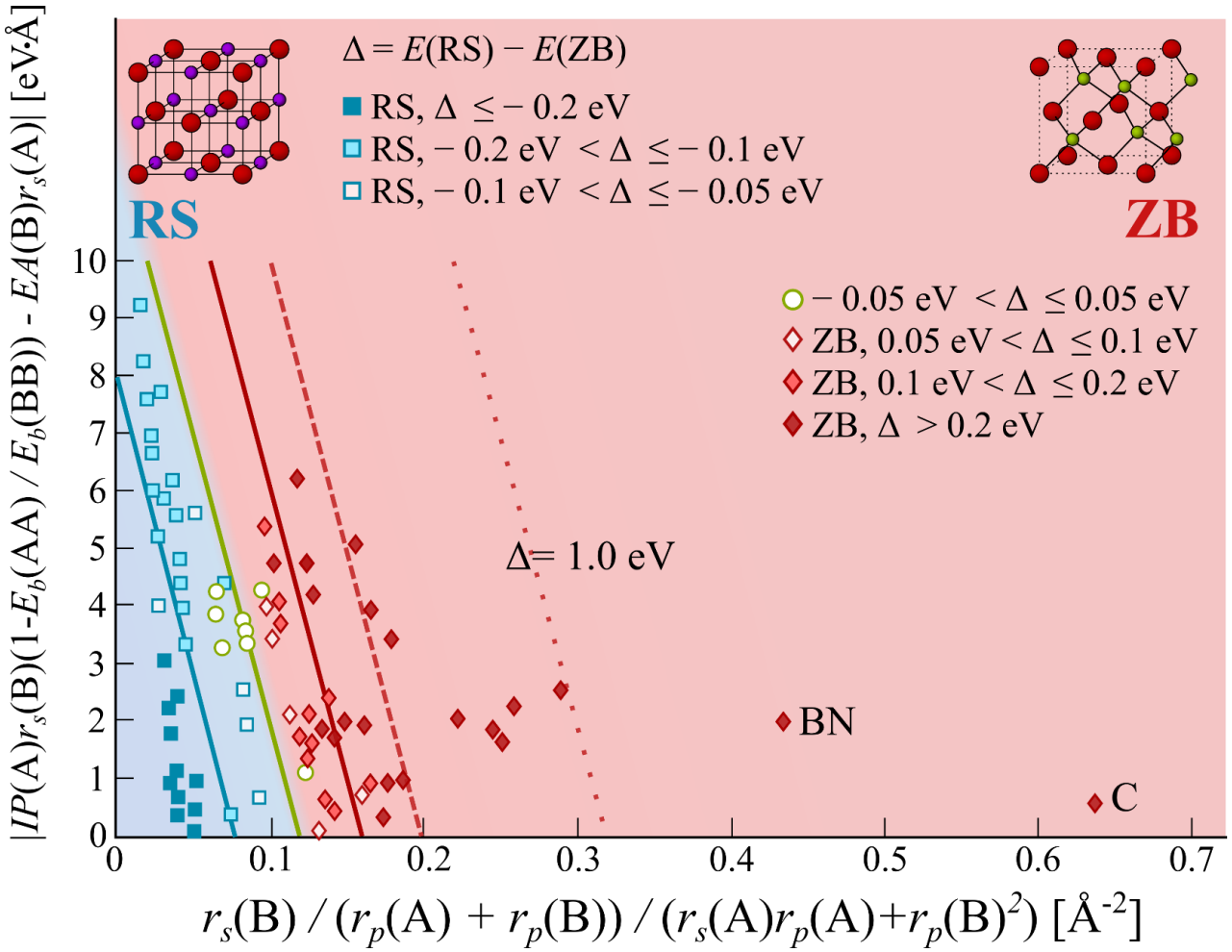
R. Ouyang, *et al.*, Physical Review Materials 2, 083802 (2018)

SISSO: Performance

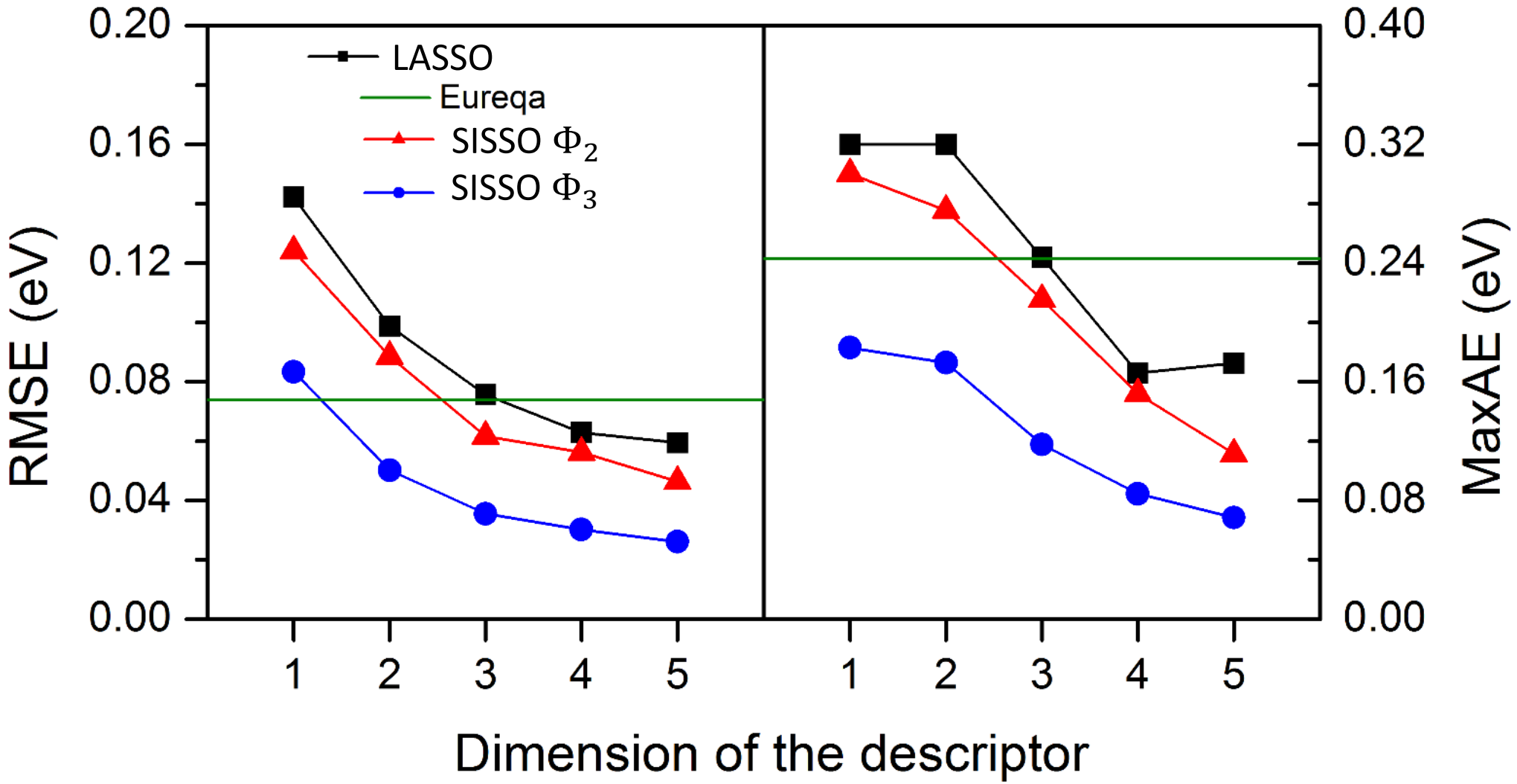
LASSO($+l_0$)



SISSO



SISSO: Performance



SISSO: Multitask and categorical

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

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

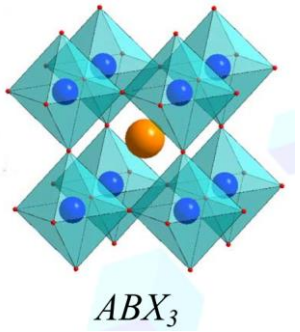
Categorical (can be also multitask): Property - material belongs to a given class (yes/no)

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

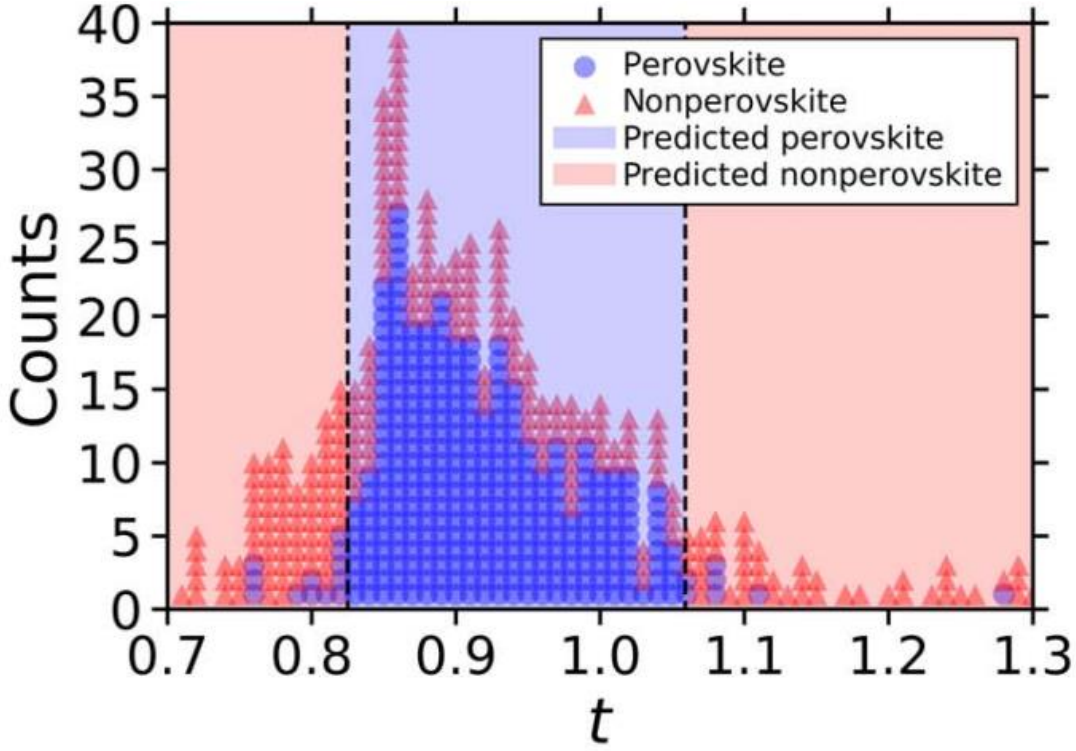
number of data in the overlap region between domains of different classes in $\boldsymbol{d}$-space

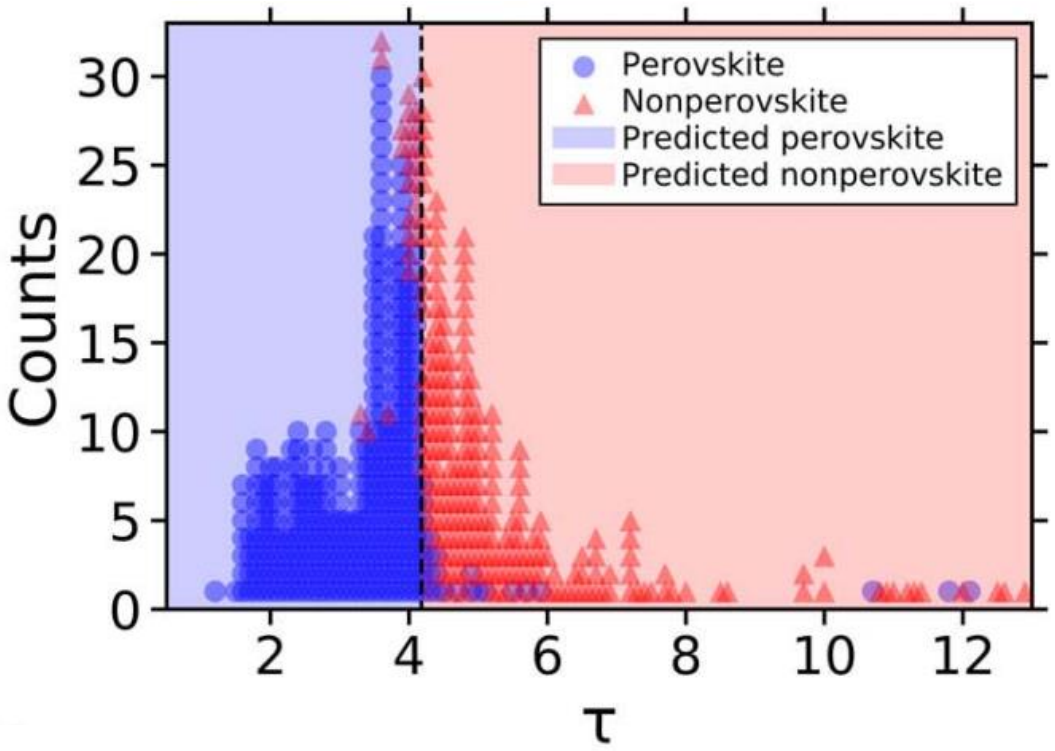
R. Ouyang, *et al*., J. Phys.: Mater. 2, 024002 (2019)

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%

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

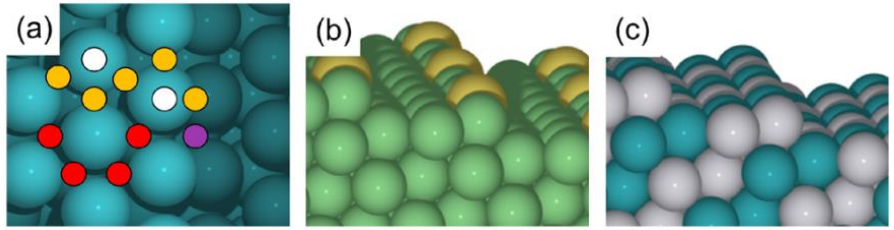
oxidation state

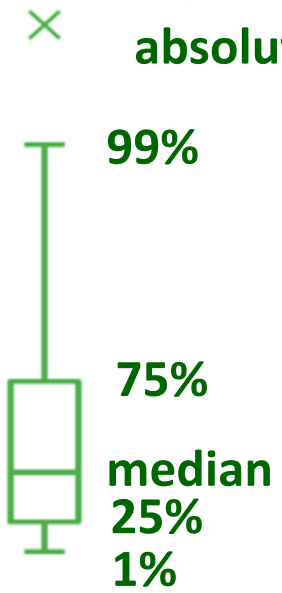
C. Bartel *et al.*, Sci. Adv. *5, eaav0693* (2019)

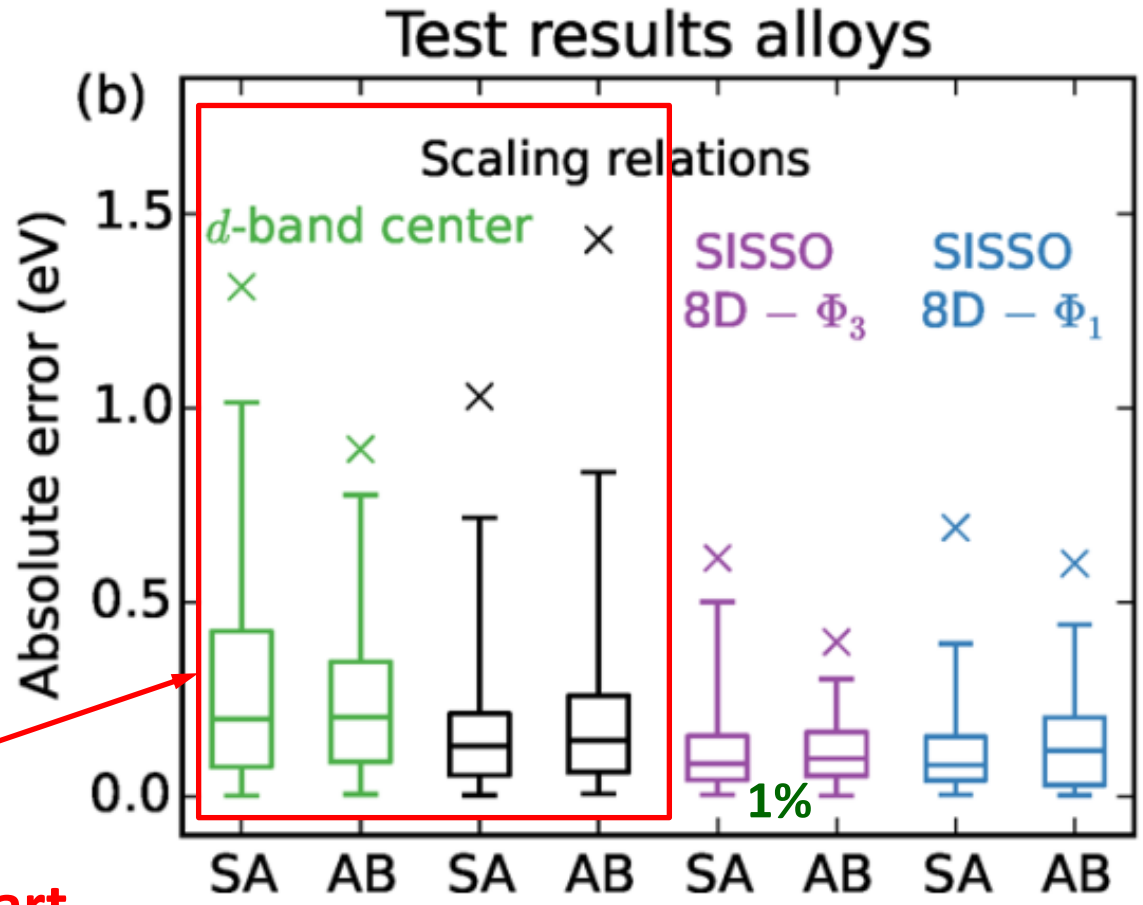
SISSO: Examples

- **Adsorption of molecules on metal surfaces**

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



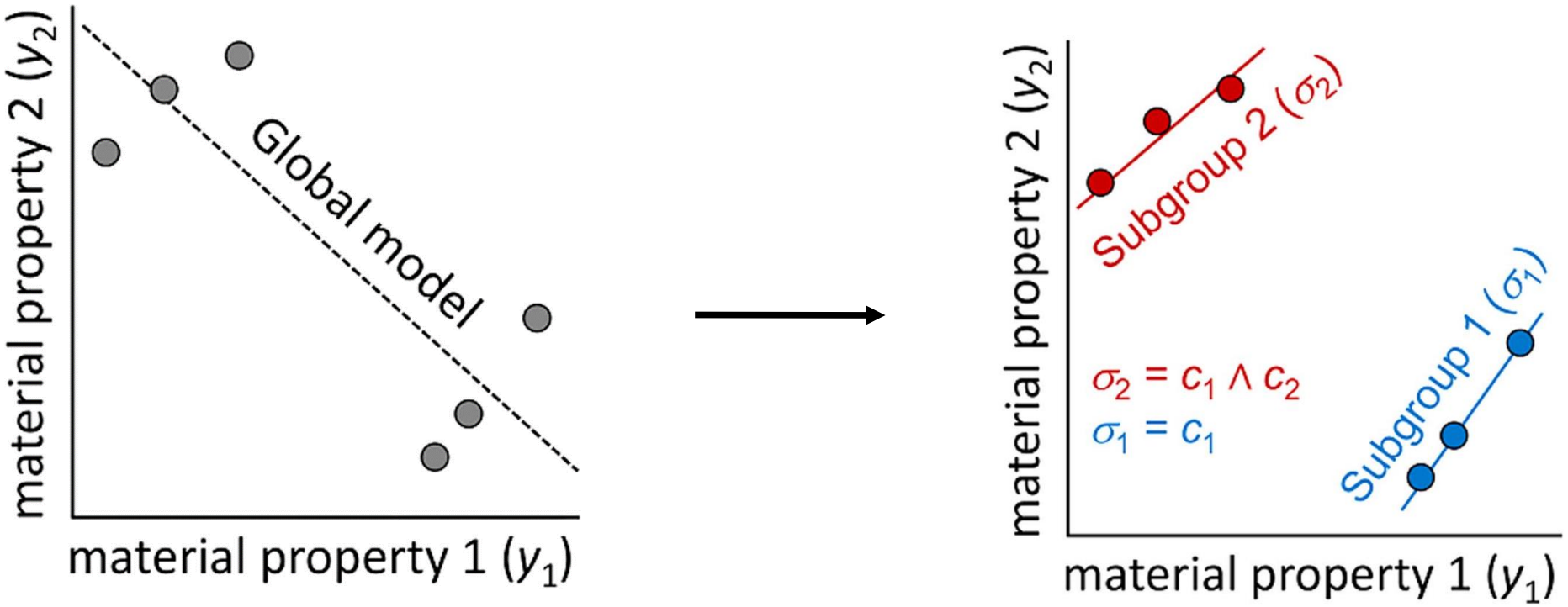




previous state of the art

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

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*

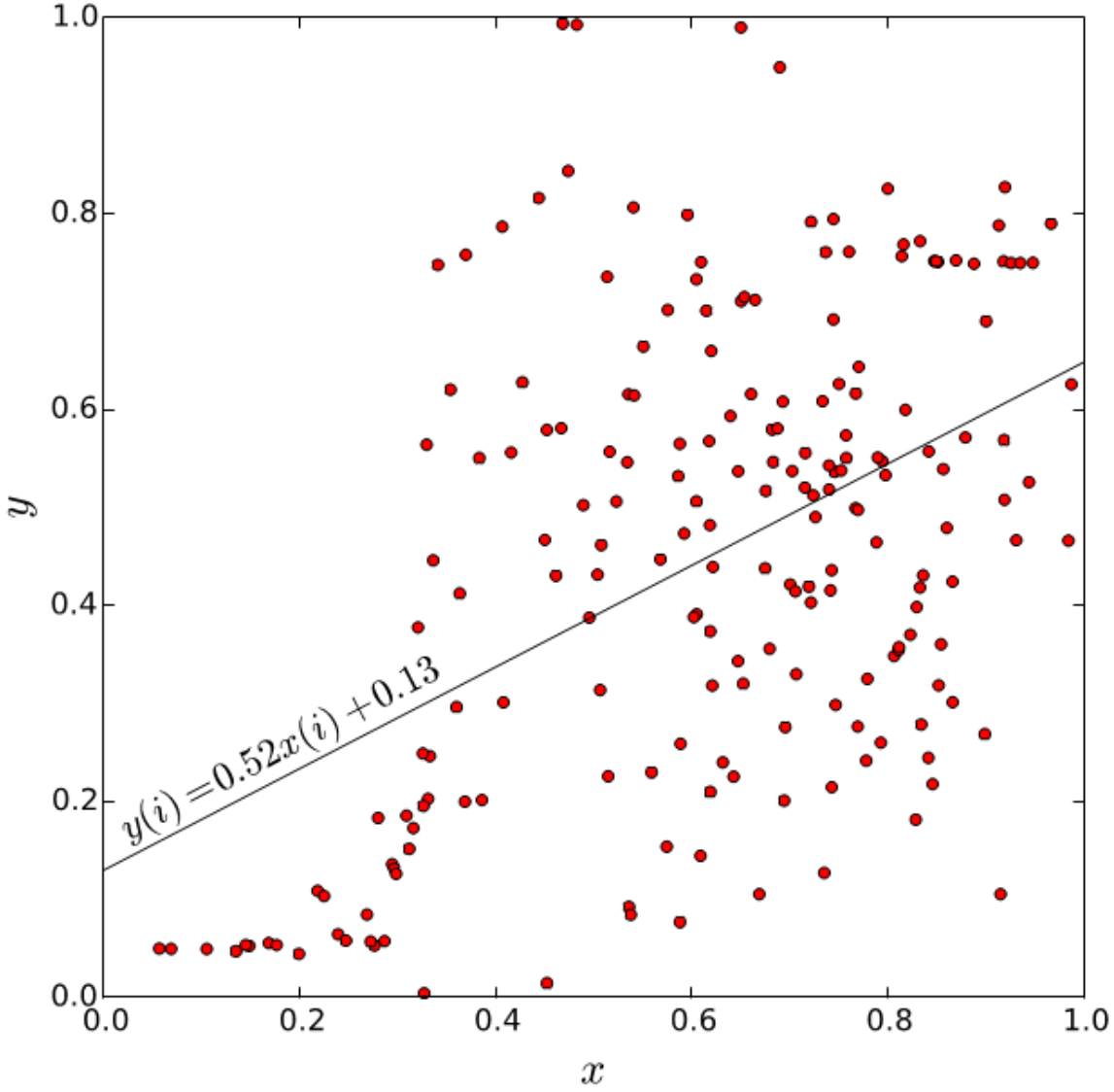
$$f = N_{subgroup}/N_{all} \times |mean_{subgroup} - mean_{all}| \times (1 - variance_{subgroup}/variance_{all})$$

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

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

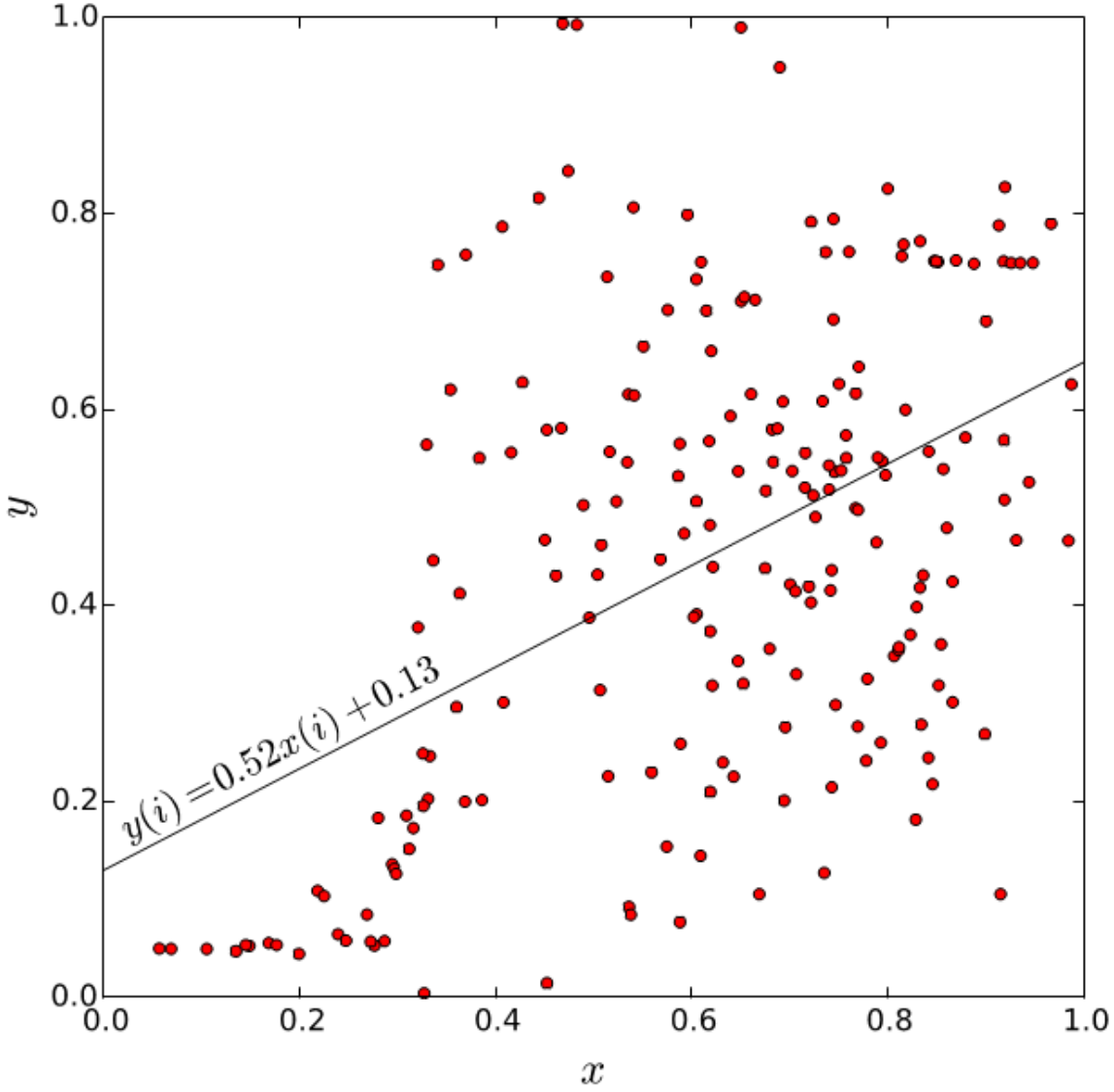
W. Klösgen, Advances in Knowledge Discovery and Data Mining. Palo Alto, CA: AAAI Press; 1996, 249

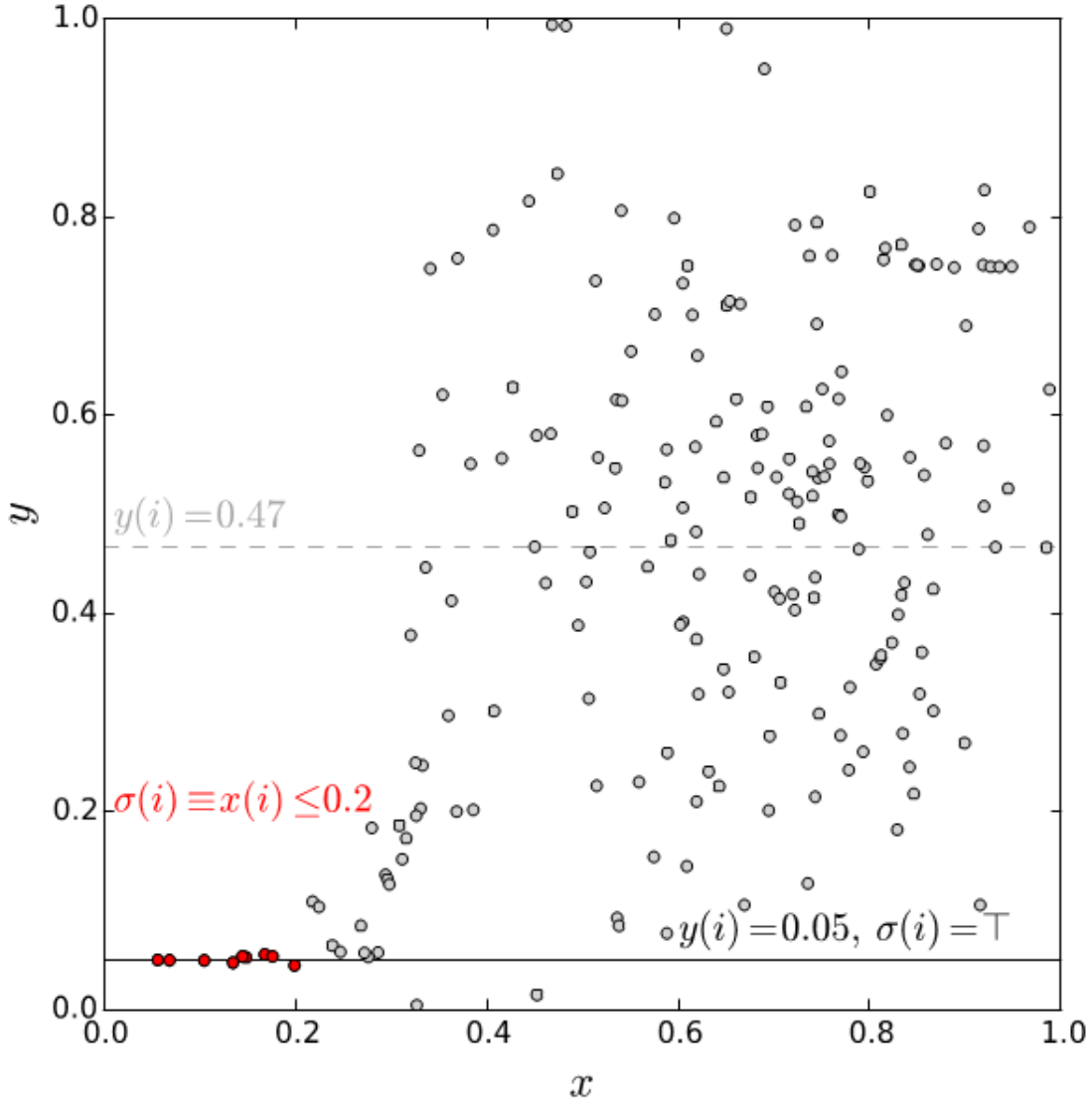
Data mining: Subgroup discovery



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

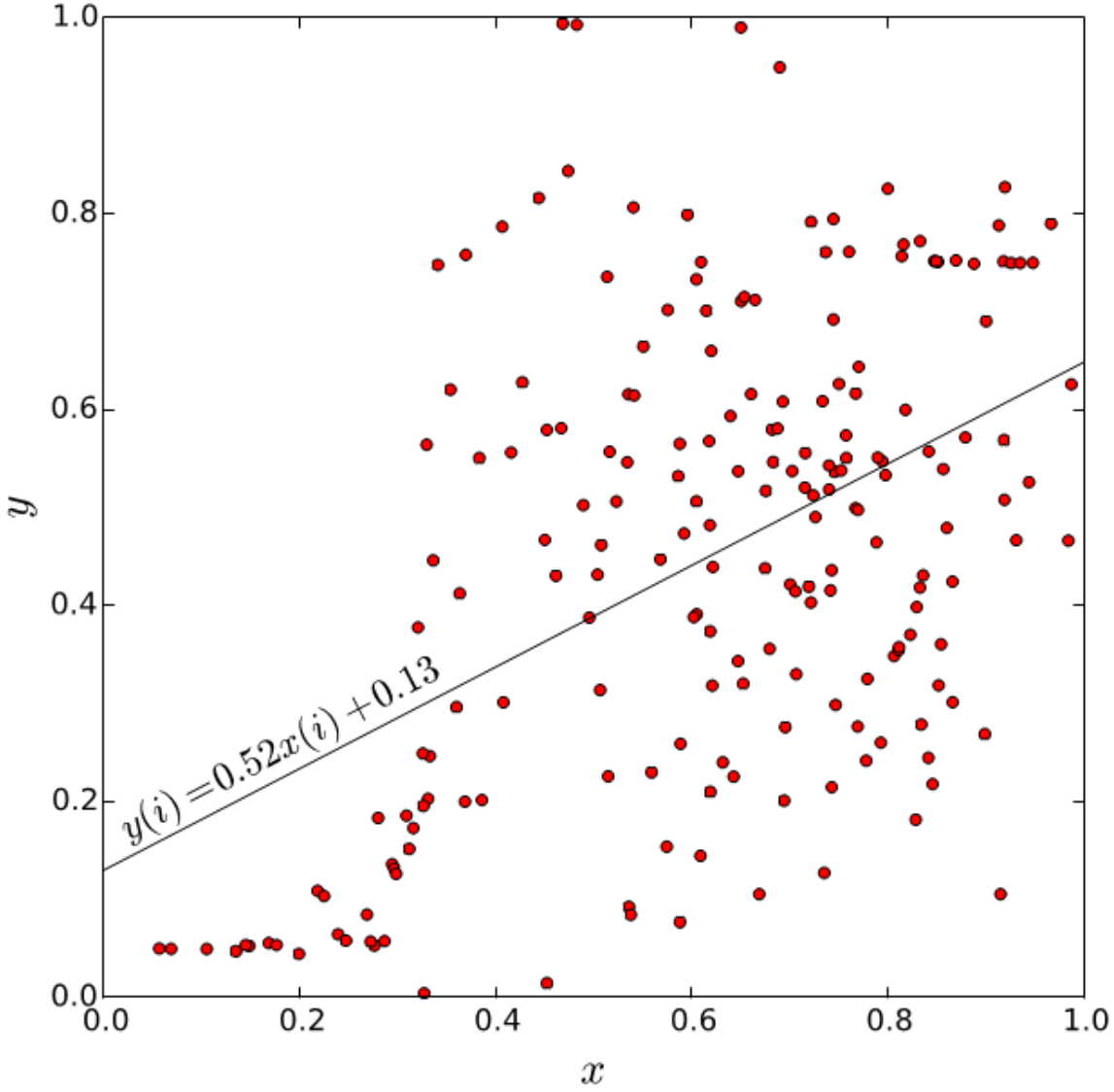
Data mining: Subgroup discovery

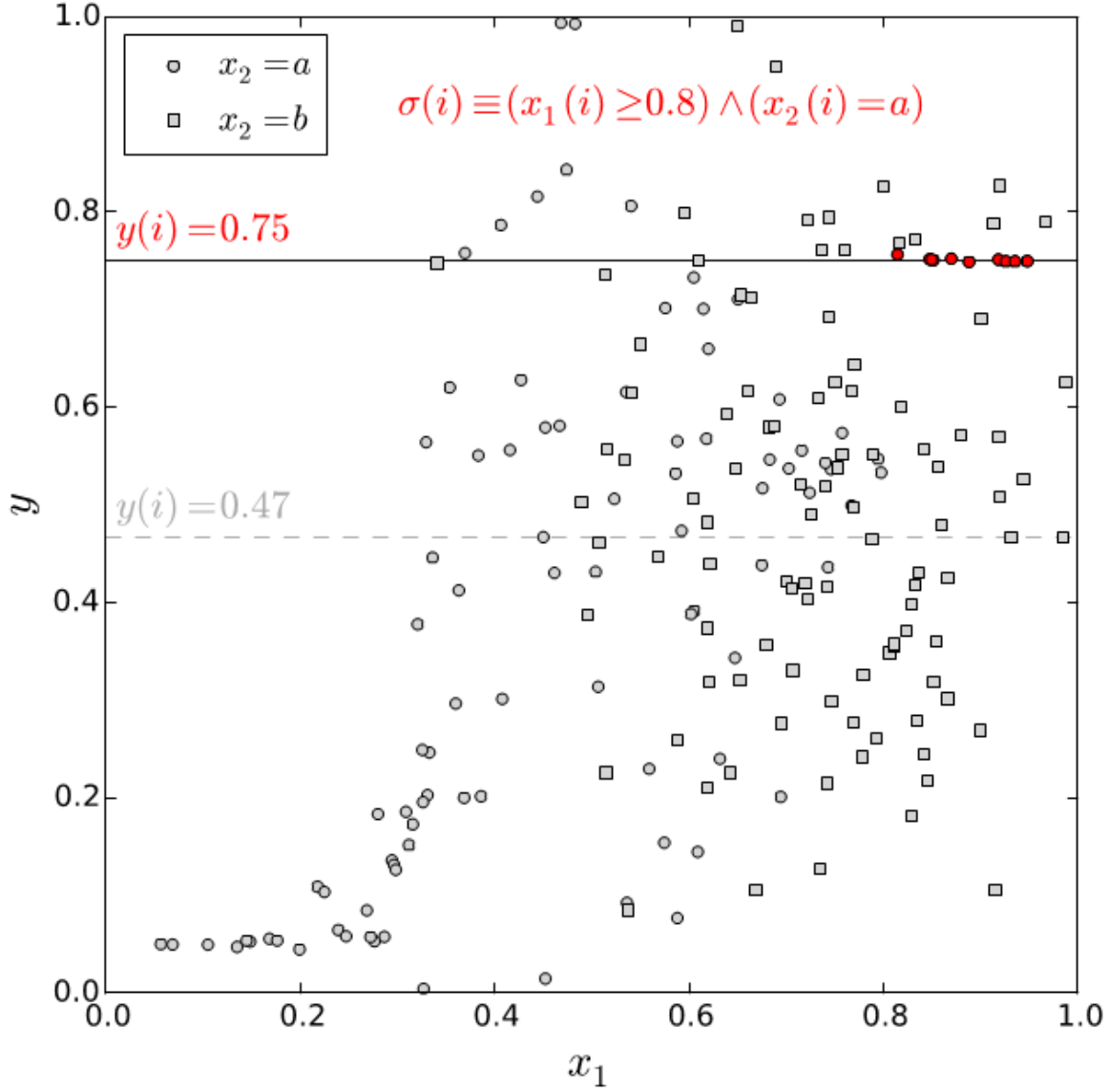




M. Boley *et al*., Data Min. Knowl. Disc. 31, 1391 (2017); B. Goldsmith *et al*., New J. Phys. 19, 013031 (2017)

Data mining: Subgroup discovery



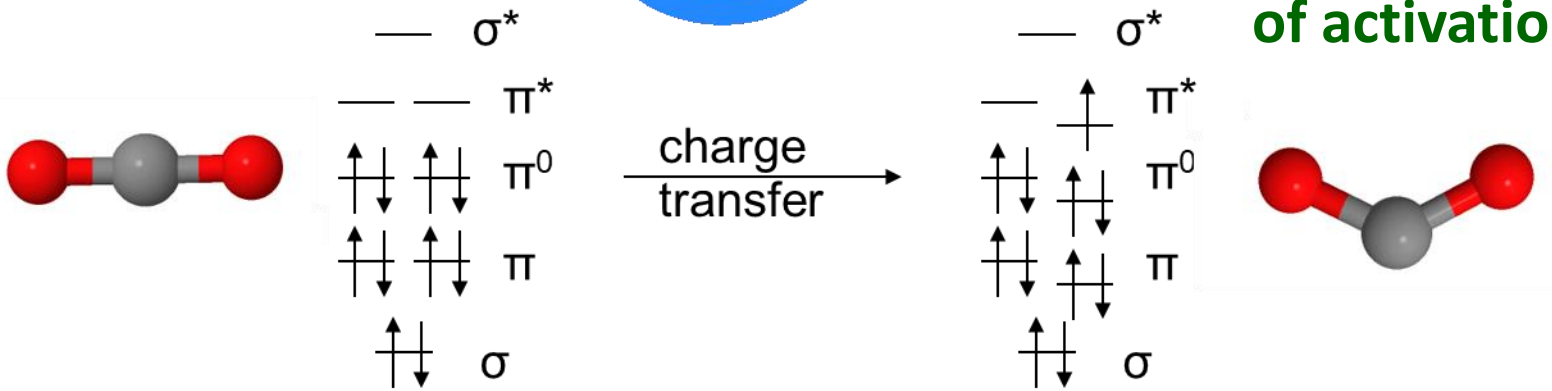


M. Boley *et al*., Data Min. Knowl. Disc. 31, 1391 (2017); B. Goldsmith *et al*., New J. Phys. 19, 013031 (2017)

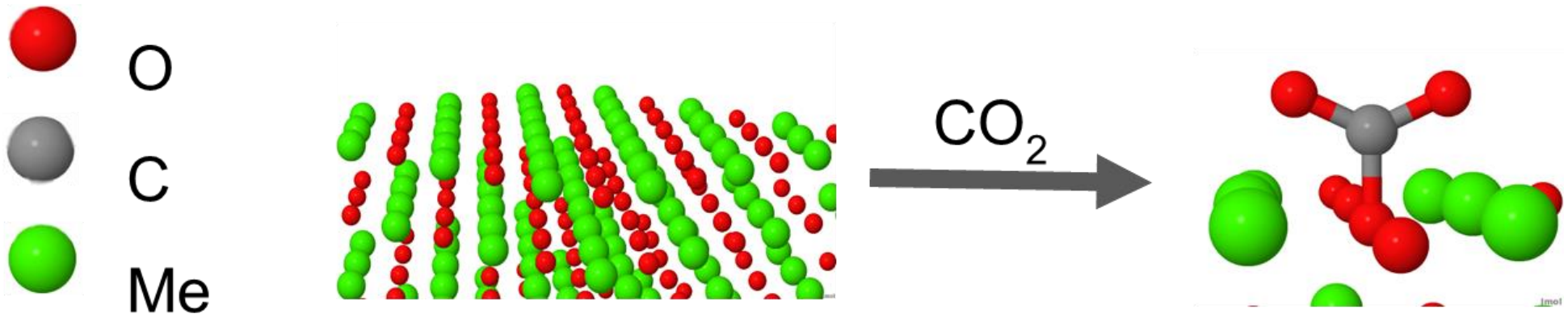
Subgroup discovery: CO_2 activation by adsorption

CO_2

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



Subgroup discovery: CO_2 activation by adsorption



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

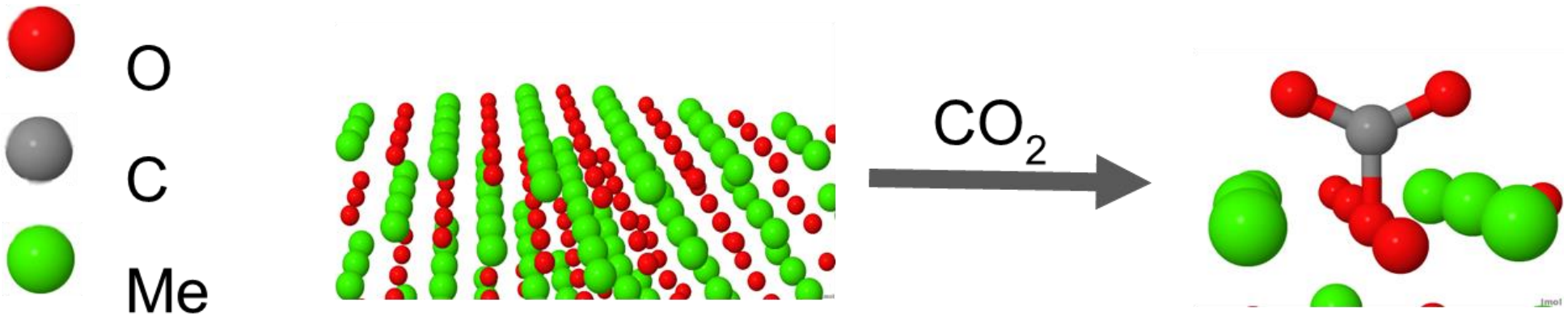
Sabatier reaction:
$CO_2 + 4H_2 = CH_4 + 2H_2O$

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

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 H 1.008																	2 He 4.0026
3 Li 6.94	4 Be 9.0122											5 B 10.81	6 C 12.011	7 N 14.007	8 O 15.999	9 F 18.998	10 Ne 20.180
11 Na 22.990	12 Mg 24.305											13 Al 26.982	14 Si 28.085	15 P 30.974	16 S 32.06	17 Cl 35.45	18 Ar 39.948
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 #	104 Rf (265)	105 Db (268)	106 Sg (271)	107 Bh (270)	108 Hs (277)	109 Mt (276)	110 Ds (281)	111 Rg (280)	112 Cn (285)	113 Nh (286)	114 Fl (289)	115 Mc (289)	116 Lv (293)	117 Ts (294)	118 Og (294)

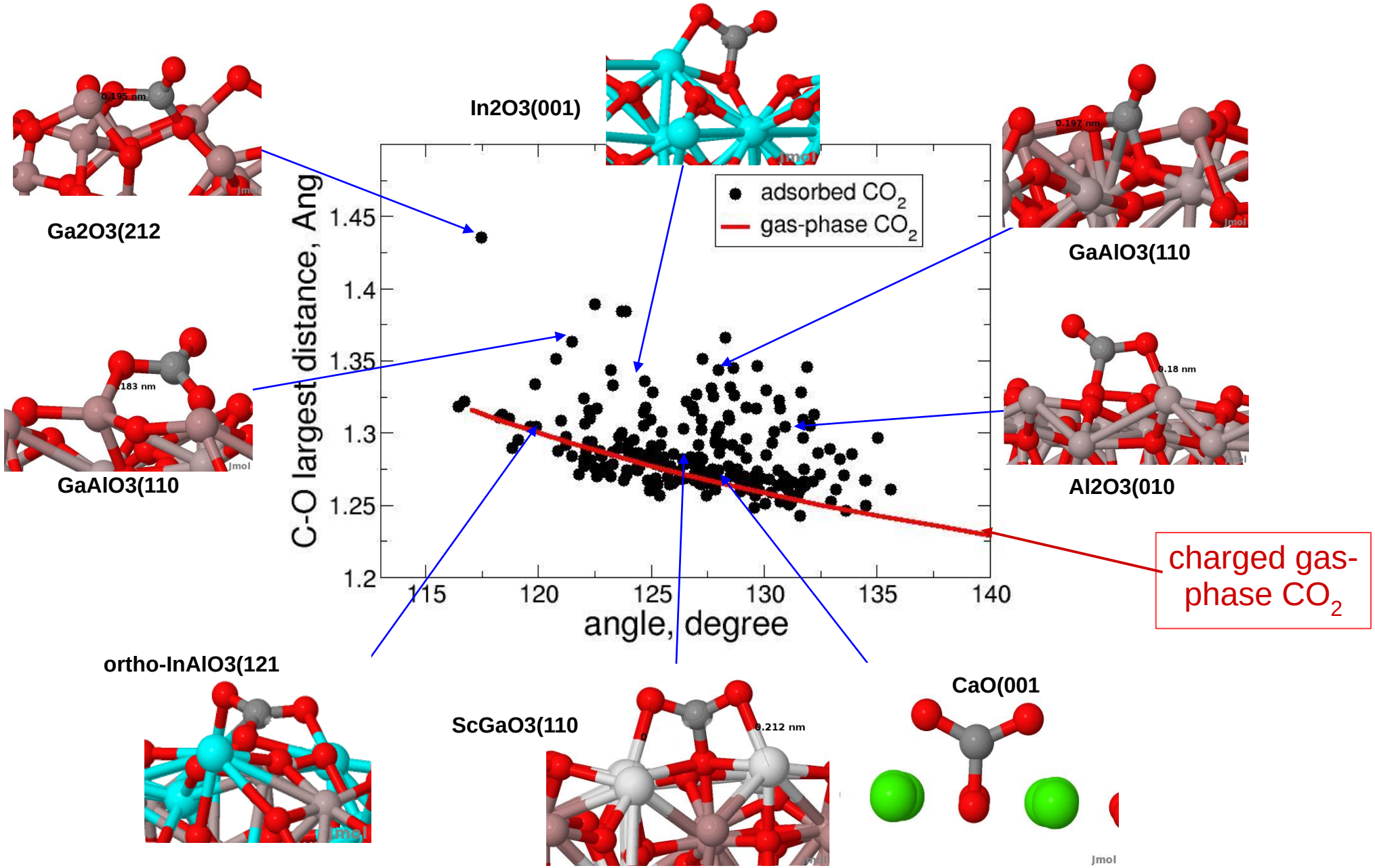
* Lanthanide series	57 La 138.91	58 Ce 140.12	59 Pr 140.91	60 Nd 144.24	61 Pm (145)	62 Sm 150.36	63 Eu 151.96	64 Gd 157.25	65 Tb 158.93	66 Dy 162.50	67 Ho 164.93	68 Er 167.26	69 Tm 168.93	70 Yb 173.05	71 Lu 174.97
# Actinide series	89 Ac (227)	90 Th 232.04	91 Pa 231.04	92 U 238.03	93 Np (237)	94 Pu (244)	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 | 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

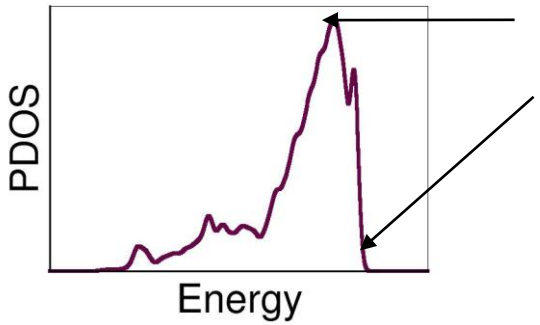
Site-specific features:

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



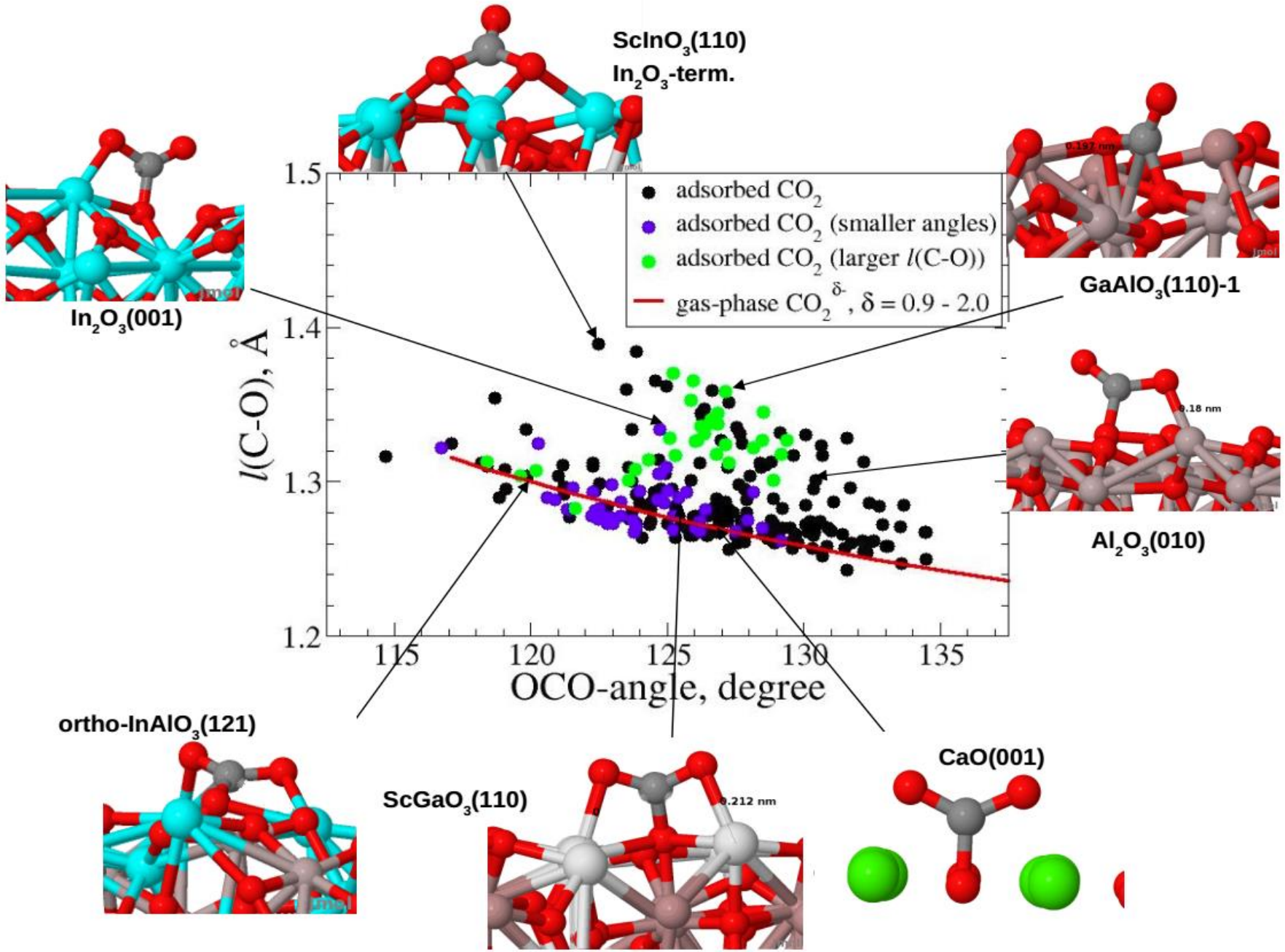
energy of maximum

energy of top

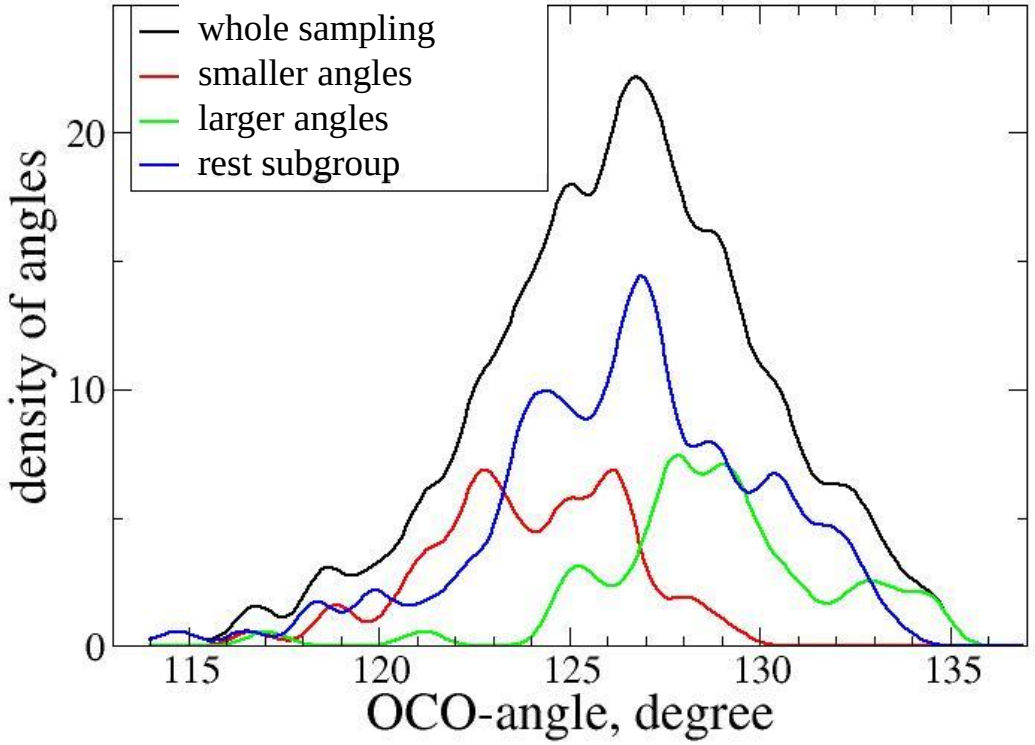
1st, 2nd, 3d, 4th moments

DOS moments: center, width, skewness, kurtosis

Subgroup discovery: Adsorbed CO2 properties

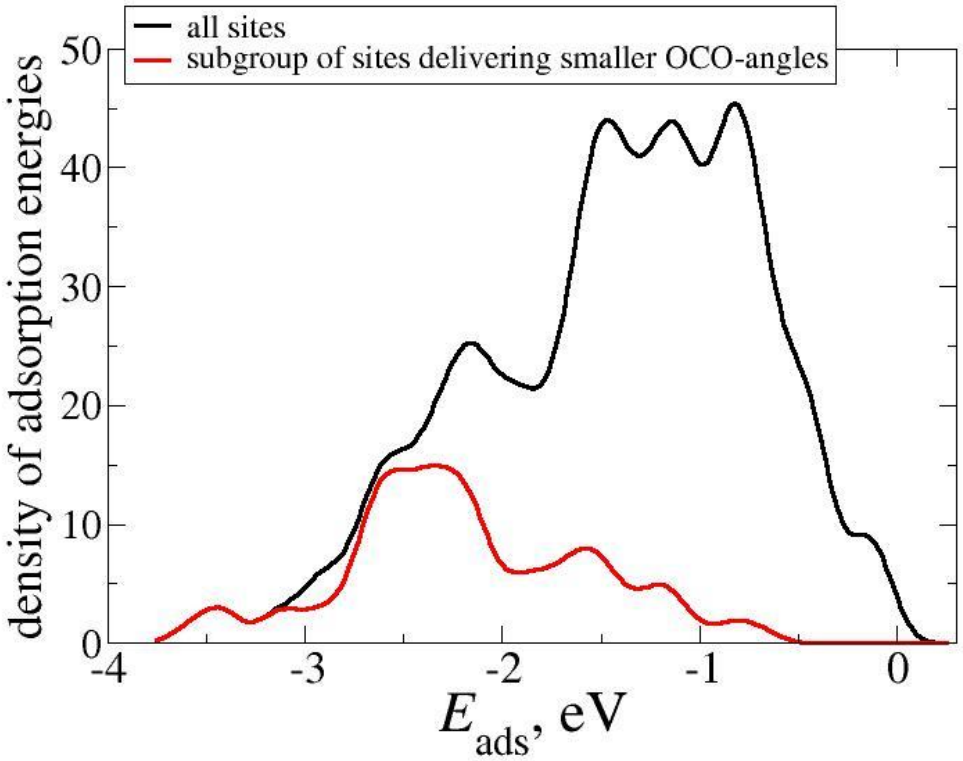


Subgroup discovery: Analysis of the OCO angle



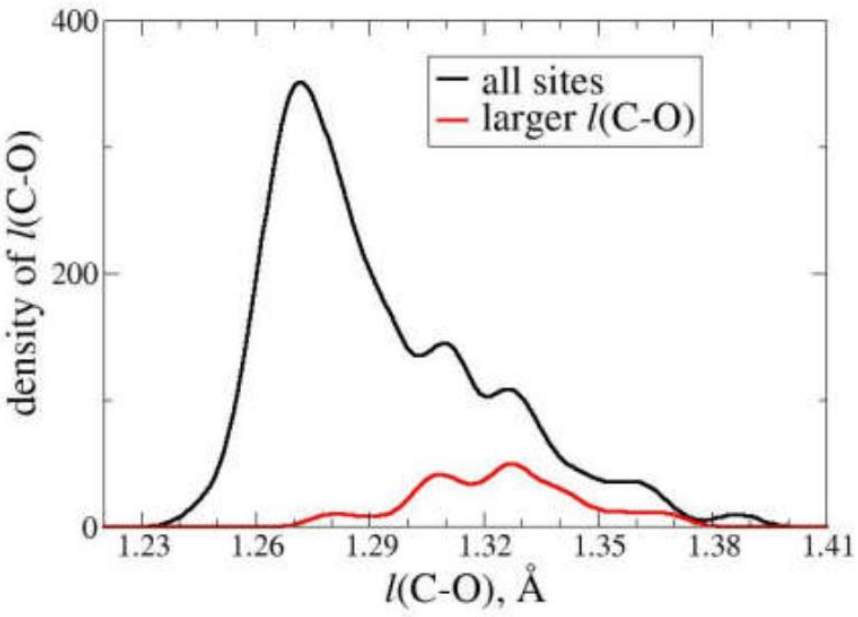
sites delivering smaller angles (59 adsorption sites):

(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)



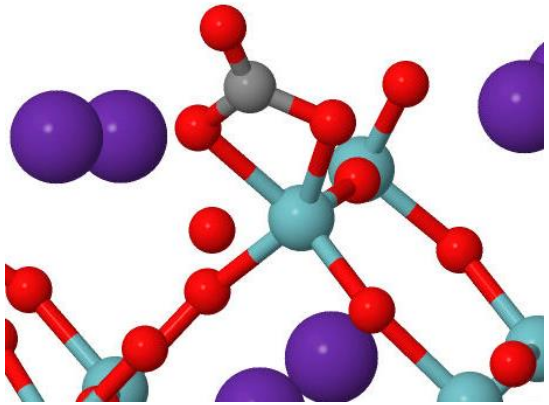
Most of the site delivering smaller OCO angles are on ionic (basic) materials

Subgroup discovery: Analysis of the C-O bond length



sites delivering larger *l*(CO) (33 sites):

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



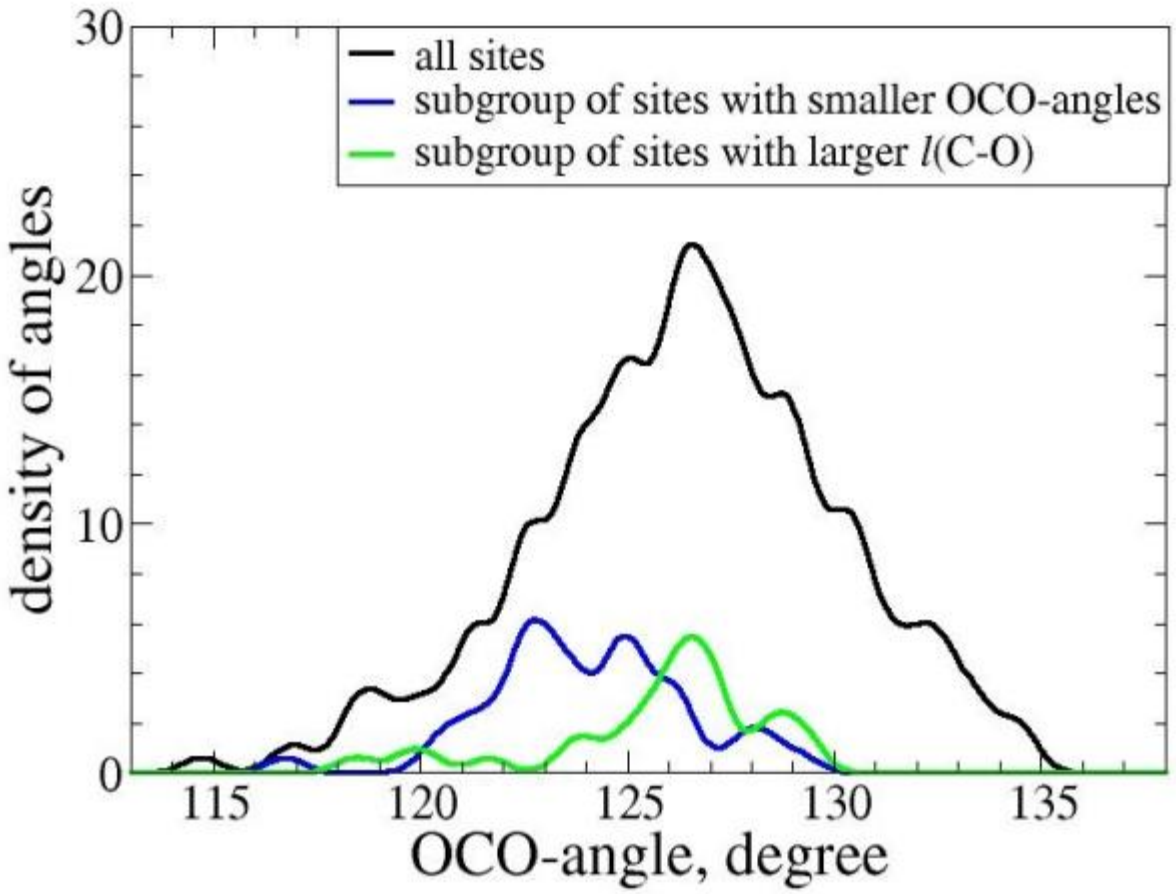
$LaGaO_3$ – cathode material in high-temperature electrochemical CO_2 reduction;

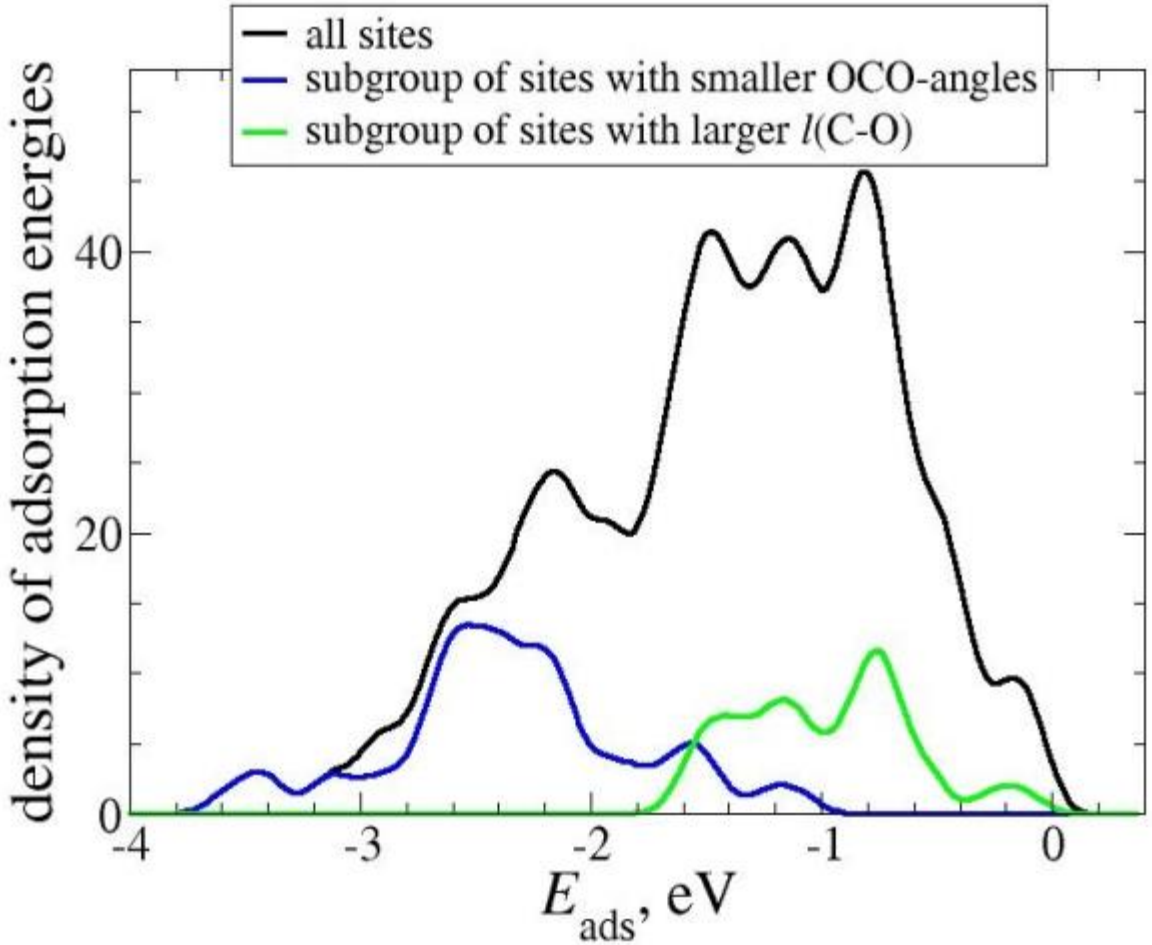
$KNbO_3$ – photocatalytic reduction of CO_2 into CH_4;

$NaNbO_3$ – 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

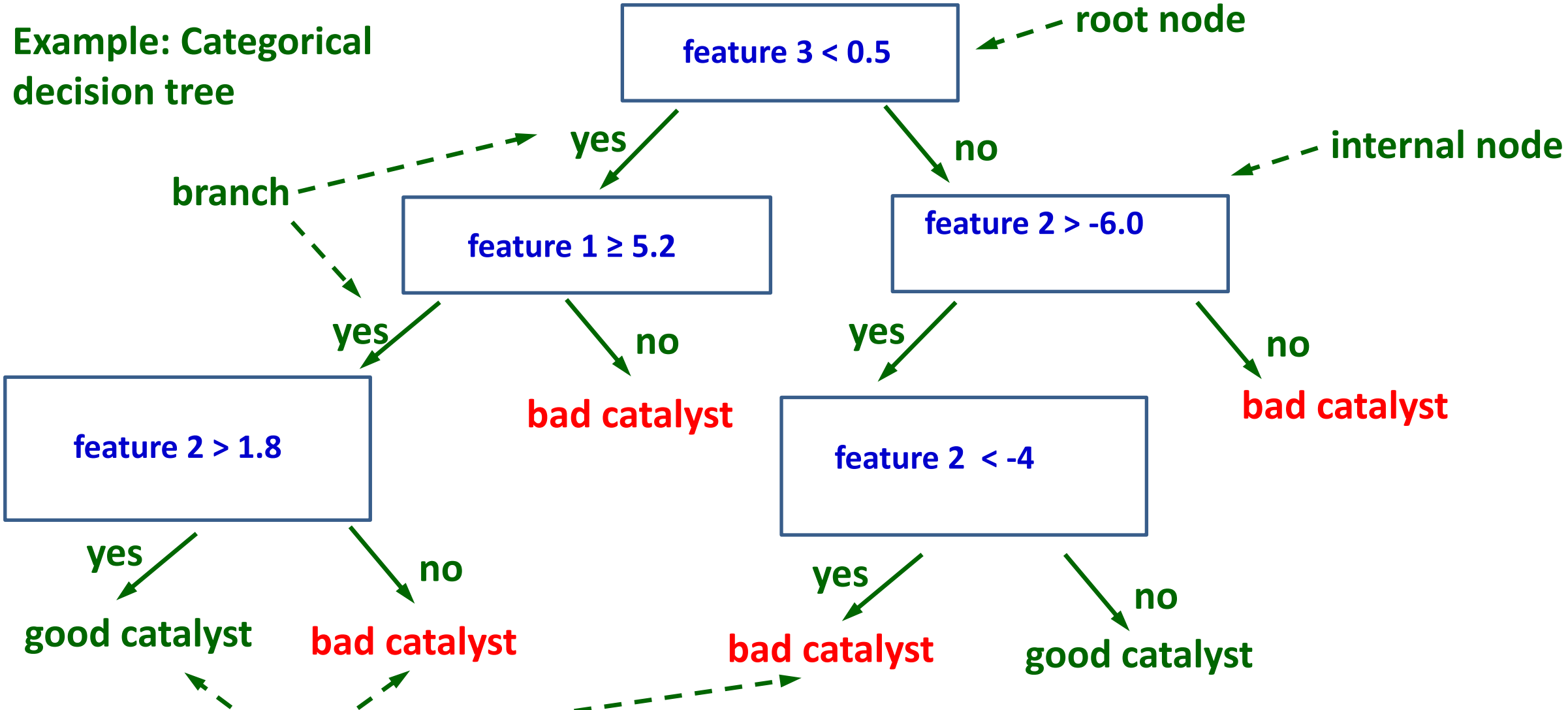
SISSO and SGD software

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

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

Decision trees

Example: Categorical decision tree

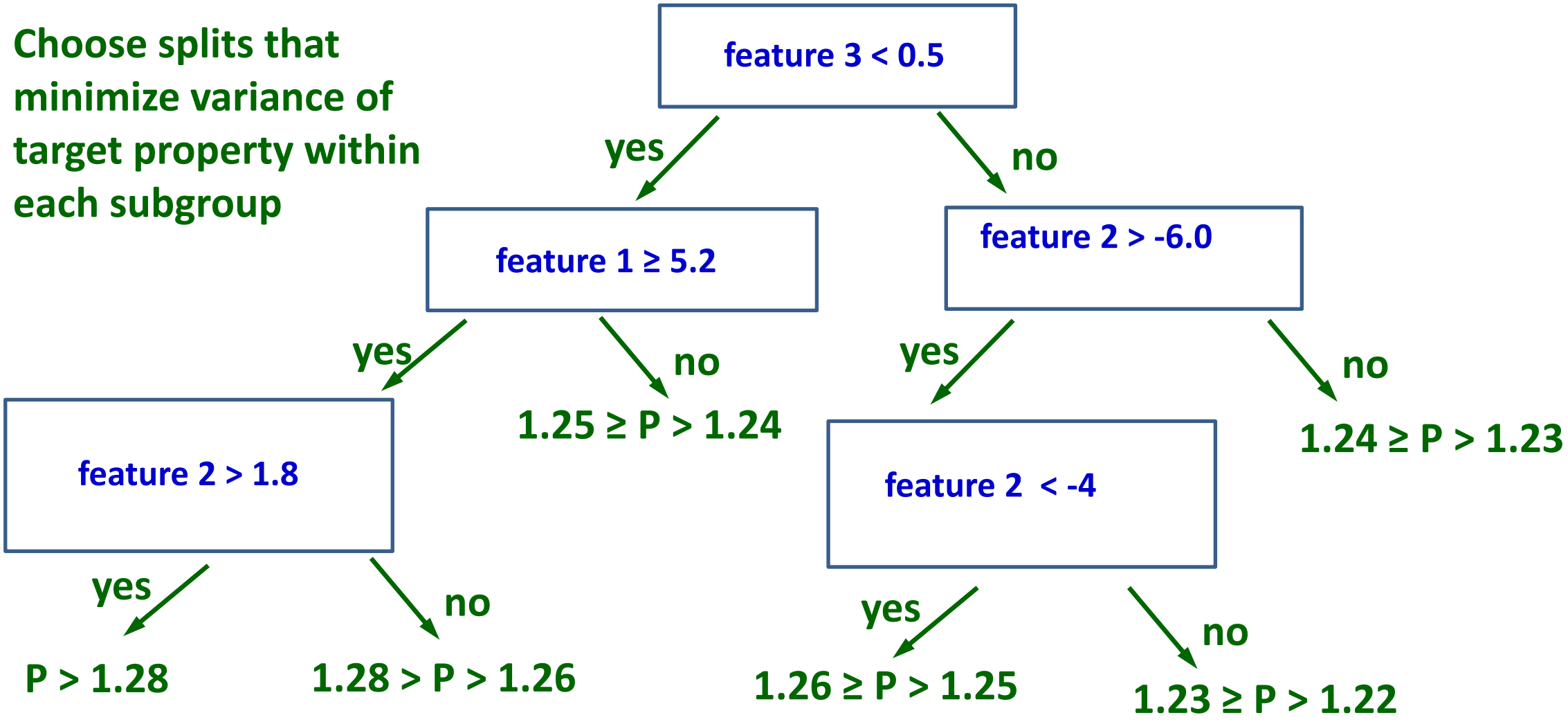


Split criterion: $\sum pk(1-pk) \rightarrow \min$

proportion of same-category inputs

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

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*,[†]

Computational databases

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

materialsproject.org

Home About Apps Documentation

The Materials Project

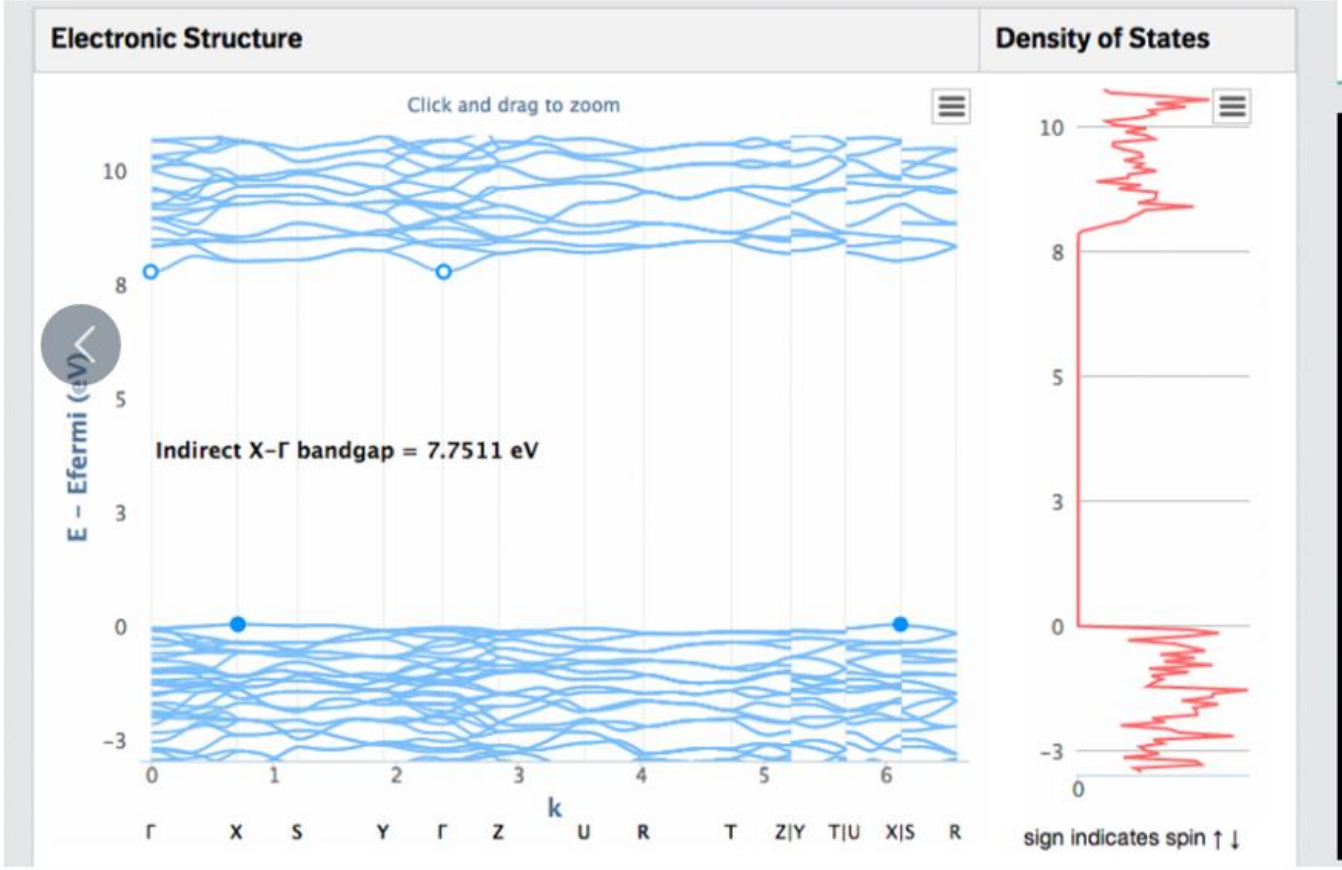
Harnessing the power of supercomputing and state of the art electronic structure methods, the Materials Project provides open web-based access to computed information on known and predicted materials as well as powerful analysis tools to inspire and design novel materials.

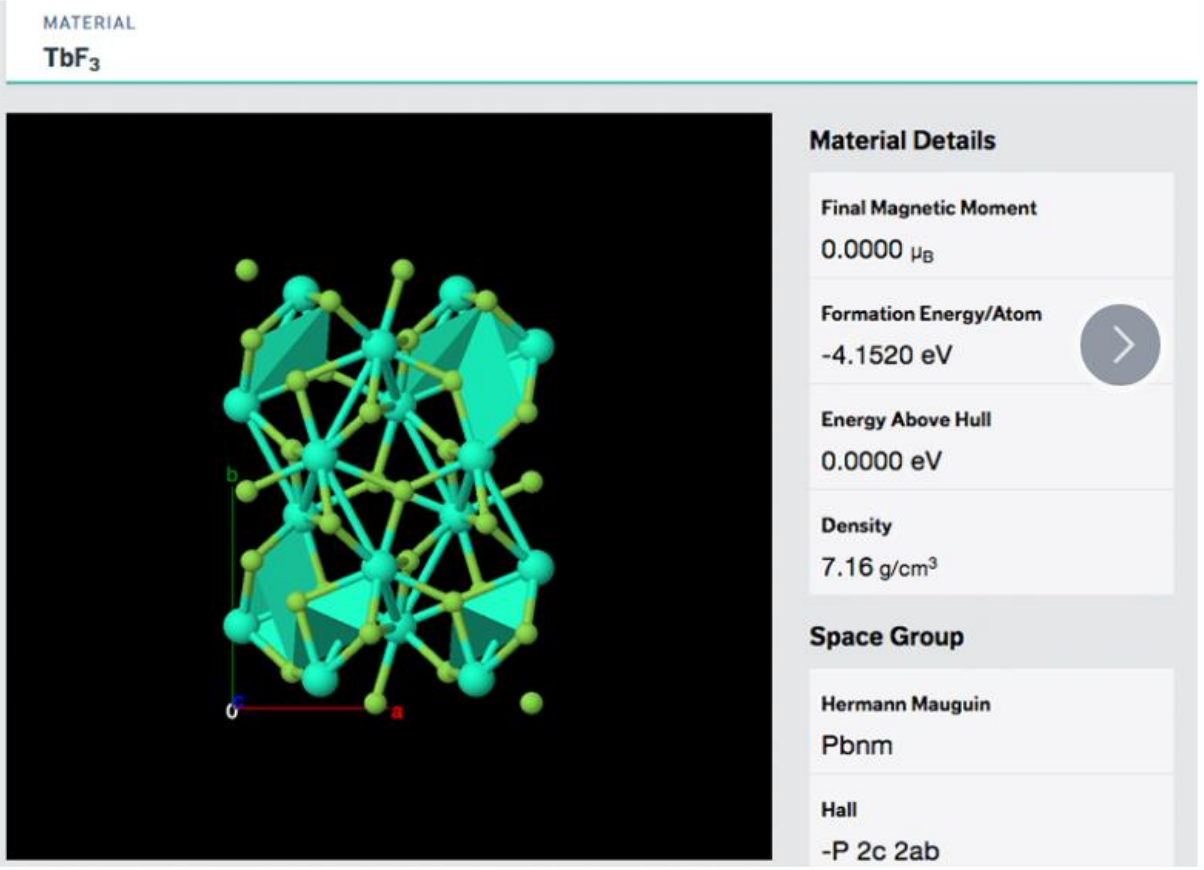
Learn more | YouTube Tutorials | Sign In or Register to start using

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/)

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	76,194	49,705	530,243
INORGANIC COMPOUNDS	BANDSTRUCTURES	MOLECULES	NANOPOROUS MATERIALS
14,071	3,411	4,730	16,128
ELASTIC TENSORS	PIEZOELECTRIC TENSORS	INTERCALATION ELECTRODES	CONVERSION ELECTRODES

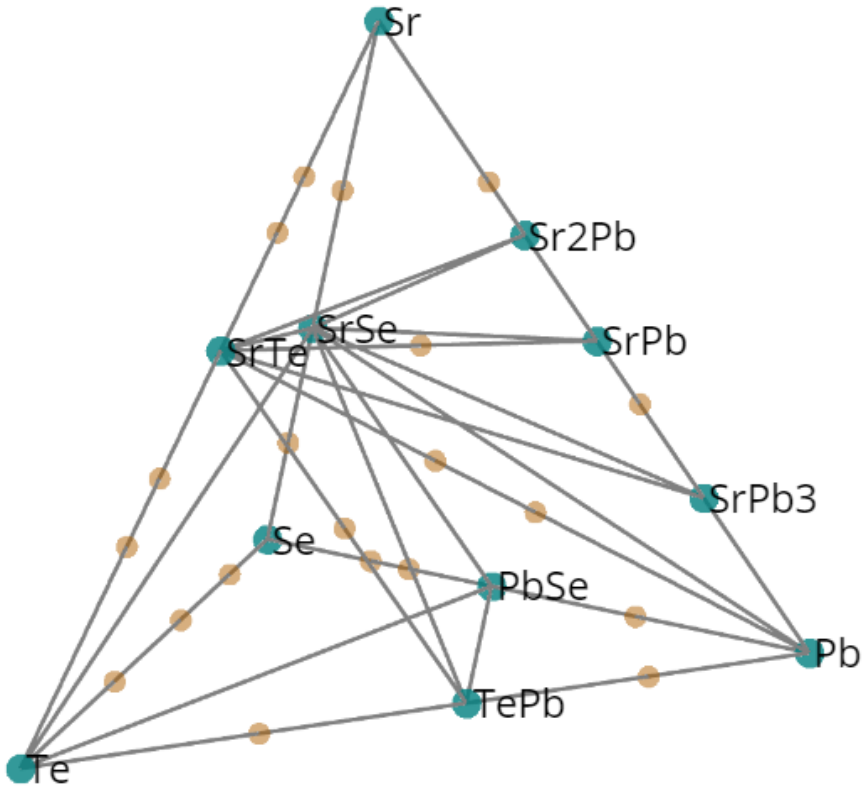
All calculations are performed with GGA or GGA+*U*

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

OQMD
The Open Quantum Materials Database

Home Materials Analysis Documentation API Download

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

All calculations are performed with GGA or GGA+*U*

Structures include also hypothetical materials (not known experimentally)

Typical data: Formation and decomposition energies

AFLOW

Automatic - FLOW for Materials Discovery

HOME | CONSORTIUM | PUBLICATIONS | FORUM | SRC | SEARCH

AFLOW SCHOOL – Online

Welcome to AFLOW, a globally available database of **3,312,125** material compounds with over **566,373,375** calculated properties, and growing.

323,516 **band structures**	125,496 **Bader charges**	6,049 **elastic properties**	6,038 **thermal properties**
1,724 **binary systems**	356,343 **binary entries**	30,071 **ternary systems**	2,400,160 **ternary entries**
	150,621 **quaternary systems**	450,567 **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)

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

NOMAD

NOVEL MATERIALS DISCOVERY

NOMAD Tutorial Series
1-2 Dec, 2020 - register now!

NOMAD Lab

Oct 13, 2020 NOMAD Tutorial 2 on Materials Encyclopedia: Registration open

REPOSITORY & ARCHIVE

MATERIALS ENCYCLOPEDIA

ARTIFICIAL INTELLIGENCE TOOLKIT

NOMAD CoE

The NOMAD (Novel Materials Discovery) Laboratory

Leader: Matthias Scheffler (Fritz Haber Institute of Max Planck Society)

Both a database and a repository (store your data)

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

AiiDA

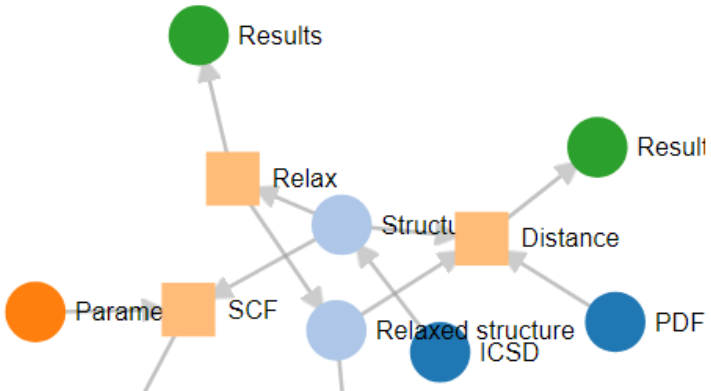
Automated Interactive Infrastructure and Database for Computational Science

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...

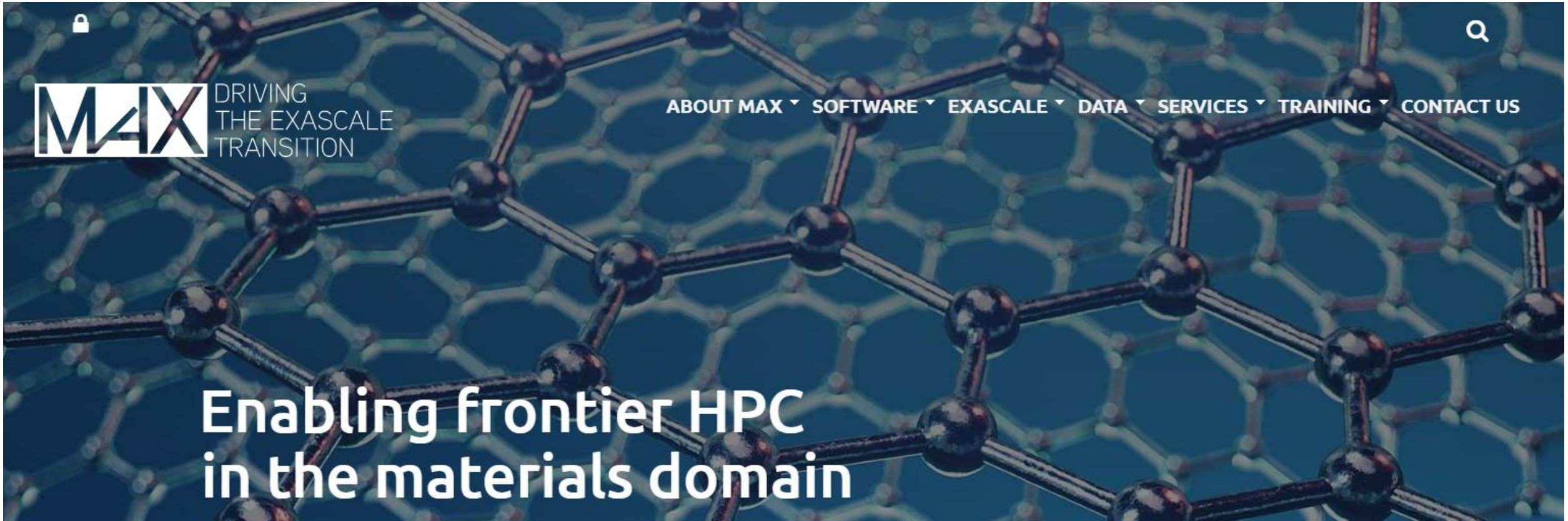
AiiDA v1.1.0 released

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)

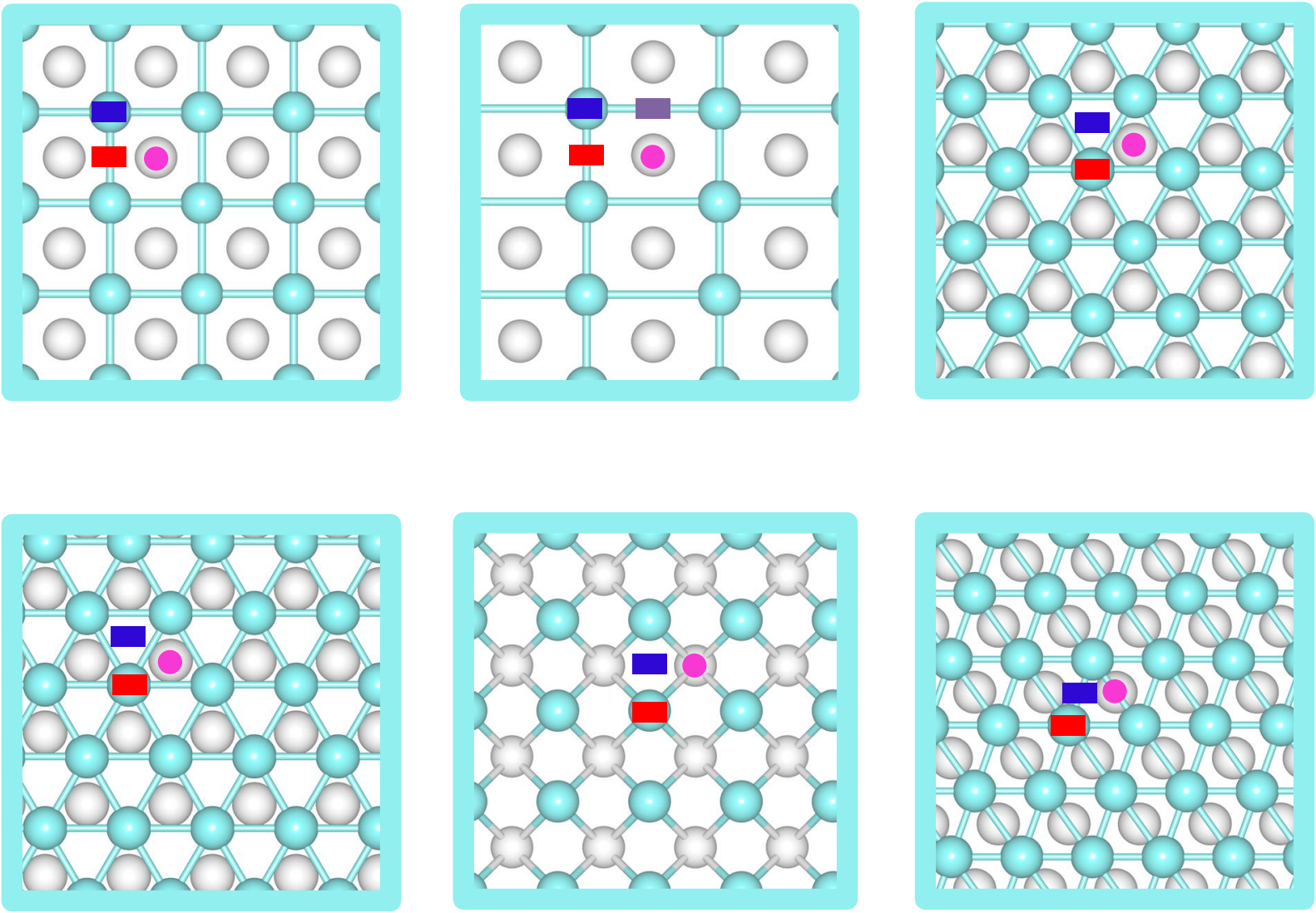
Leader: Nicola Marzari (EPFL, Switzerland)

Provides tools for performing high-throughput calculations



SISSO tutorial

Example: Water molecule adsorption energy on metal surfaces: *d*-band center versus SISSO



Training data:
45 different transition metal surfaces
adsorption energies of the most stable adsorption configurations
(totally 45 data points)

SISSO tutorial: Primary features

Class	Name	Abbreviation
Atomic	Atom radius	R
	Electronegativity	E
	HOMO	H
	LUMO	L
	Ionization energy	I
Bulk	*d* band center	DB
	Fermi energy	F
Surface	*d* band center	DS
	Chemical potential	C
	Coordination number	CN
	Effective coordination number	ECN