Finding patterns, correlations, and descriptors in materials data using subgroup discovery and compressed sensing

Bryan R. Goldsmith
University of Michigan, Ann Arbor
Department of Chemical Engineering

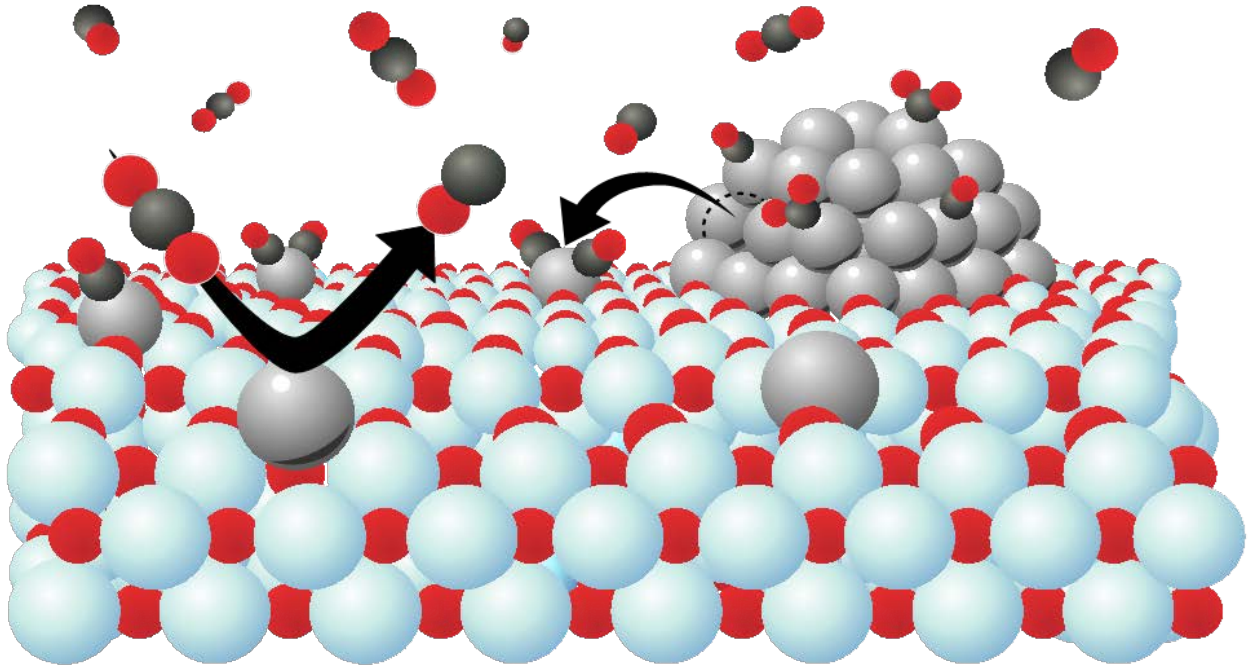
Christopher J. Bartel and Charles Musgrave
CU Boulder, Department of Chemical and Biological Engineering

Chris Sutton, Runhai Ouyang, Luca M. Ghiringhelli, Matthias Scheffler
Fritz Haber Institute of the Max Planck Society, Theory Department

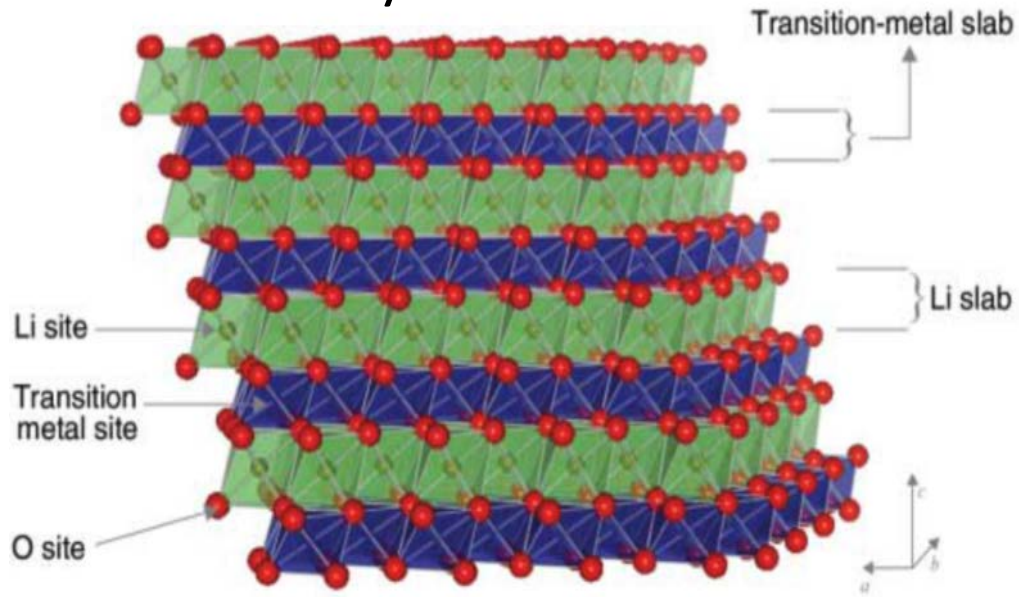
Mario Boley and Jilles Vreeken
Max Planck Institute for Informatics

Predicting advanced materials requires understanding the mechanisms underlying their function

Catalysts



Battery materials



[1]

[1] K. Kang *et al. Science* 311, 977 (2006)

Identifying physically meaningful *descriptors* can aid materials discovery

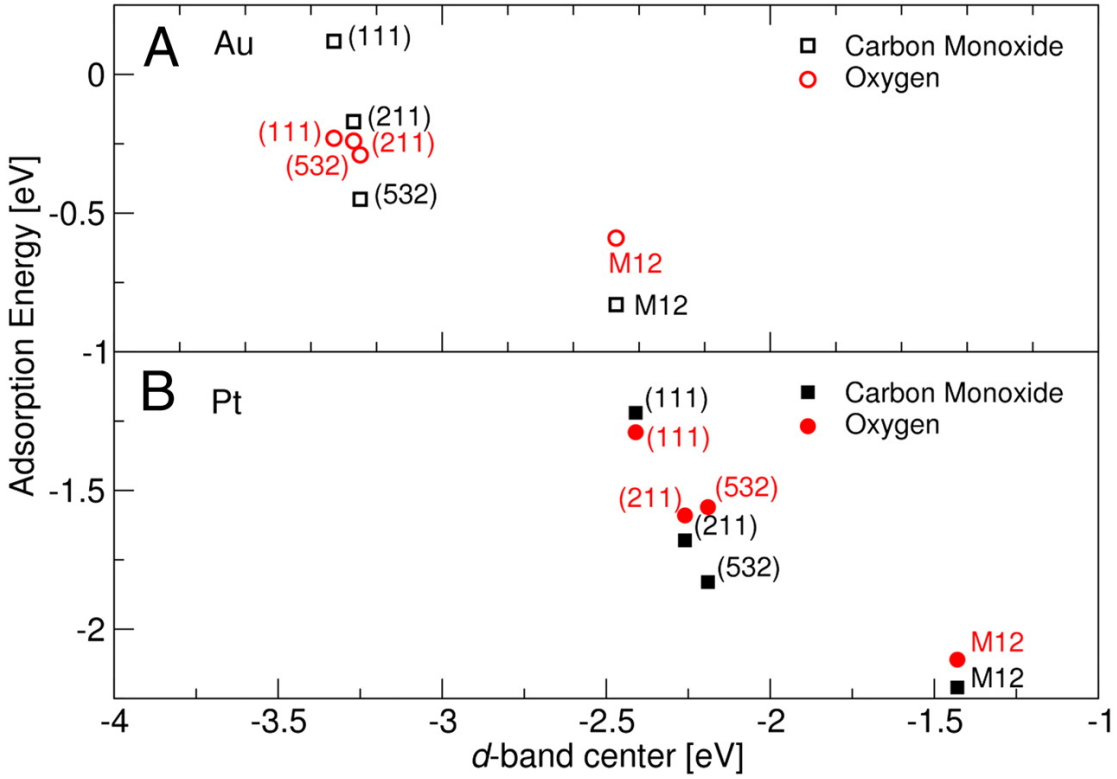
Descriptor → Property

Descriptor = function(atomic or material features)

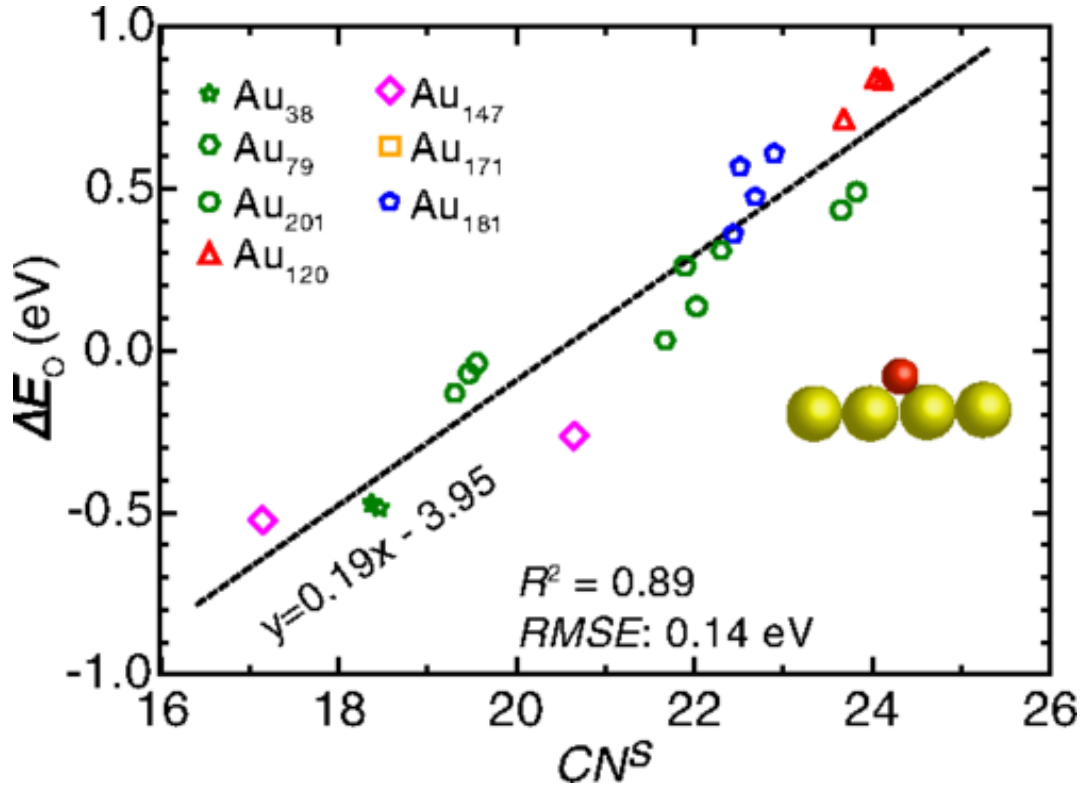
Screen new catalysts

Increase understanding

Nørskov, Jens K., *et al. PNAS* 108.3 (2011) 937-943.



Ma, Xianfeng and Hongliang Xin *PRL* 118.3 (2017) 036101.



Identifying physically meaningful *descriptors* can aid materials discovery

Descriptor → Property

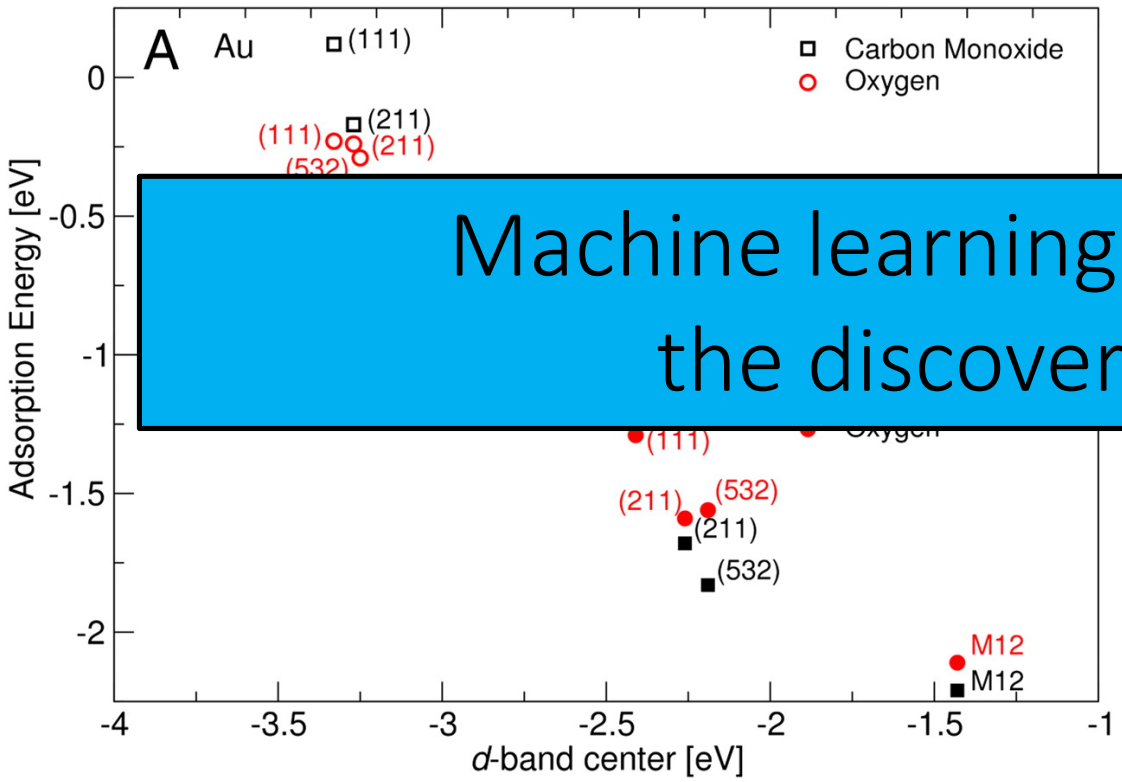
Descriptor = function(atomic or material features)

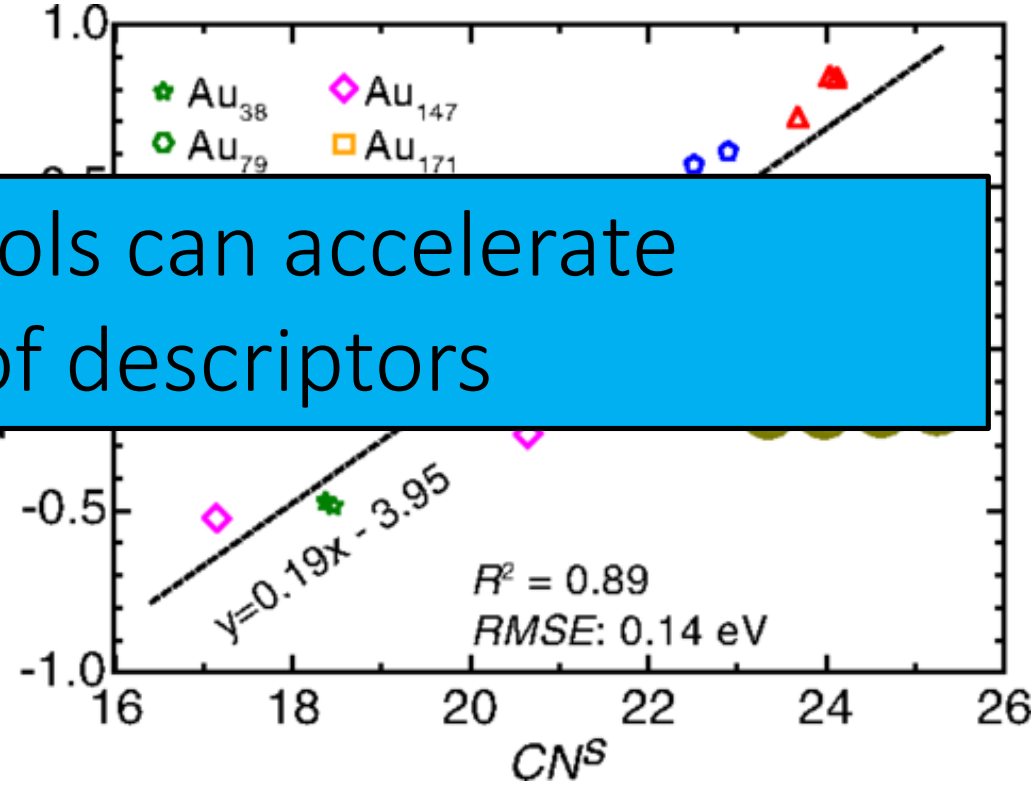
Screen new catalysts

Increase understanding

Nørskov, Jens K., *et al. PNAS* 108.3 (2011) 937-943.

Ma, Xianfeng and Hongliang Xin *PRL* 118.3 (2017) 036101.





Machine learning tools can accelerate the discovery of descriptors

This talk focuses on two data-analytics tools to find descriptors of materials

1. ***Compressed sensing*** **to find low-dimensional descriptors**
   - Perovskite oxides and halides

2. ***Subgroup discovery*** **to find local patterns and their descriptions**
   - Gold clusters in the gas phase (sizes 5-14 atoms)
   - Octet binary (AB) semiconductors

3. **Future work: Compressed sensing and subgroup discovery for catalysis**

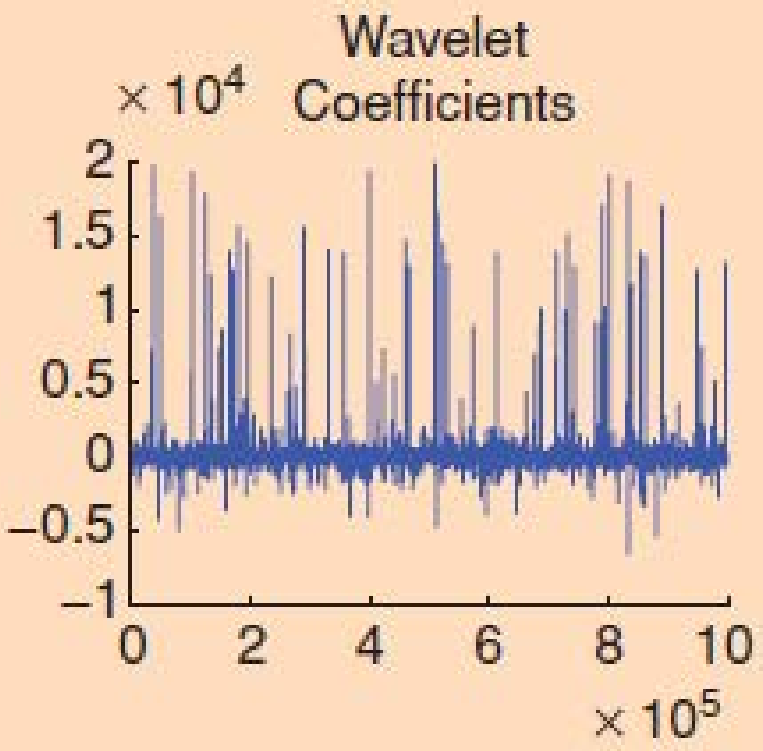
Part 1. Compressed sensing to find interpretable descriptors

Compressed sensing allows the construction of sparse models with high accuracy

Original image

Sparse in the basis set

Recovered with 10% measurements



Emmanuel Candès, Terence Tao, and David Donoho

Compressed sensing allows construction of sparse models with high accuracy

Ideally use l_0-norm minimization

$$\min \|\beta\|_0 \text{ subject to } y = \boldsymbol{D}\beta$$

coefficients

l_0-norm: total # of non-zero coefficients

Target property

Matrix of the materials' features

Emmanuel Candès, Terence Tao, and David Donoho

l_0-norm minimization is too expensive to perform for large feature matrix *D*!

Instead often minimize l_1-norm (LASSO) as approximation of l_0-norm

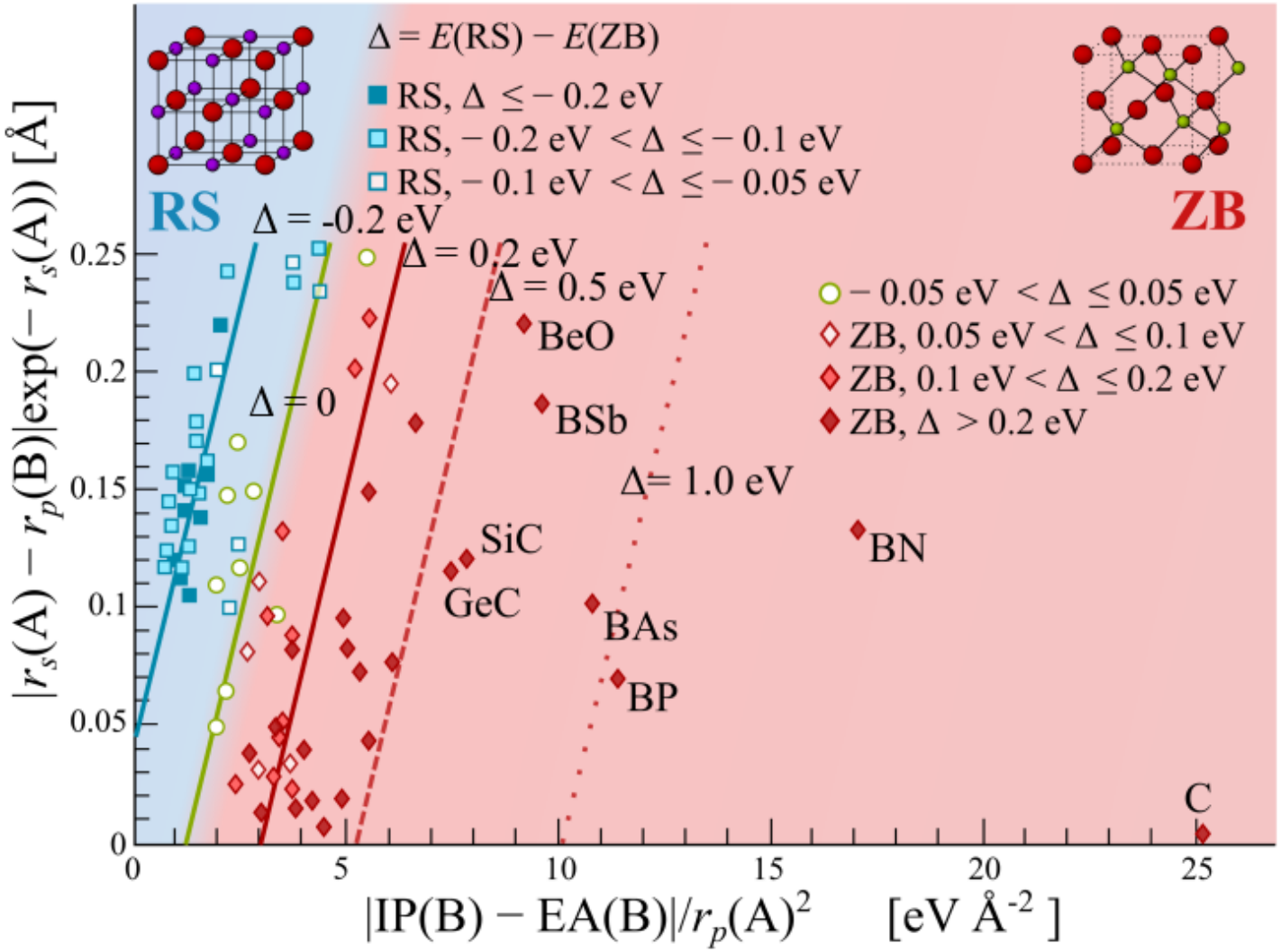
l_1-norm:
Sum of absolute value of coefficients

$$\hat{\beta}_{LASSO}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{2} \|y - \boldsymbol{D}\beta\|_2^2 + \lambda \|\beta\|_1 \right)$$

Root mean squared error

Regularization parameter

Example of LASSO+l_0: Find a descriptor that predicts the crystal structure energy differences between the 82 octet AB compounds

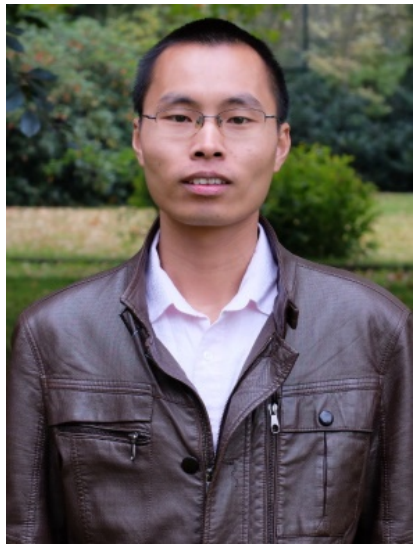


LASSO+l_0: L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, *PRL 114*, 105503 (2015)

Managing high dimensional and correlated feature spaces by combining screening and compressed sensing

Unfortunately LASSO has stability issues for a huge feature space of correlated features

→ This issue has been solved recently using the Sure Independence Screening Sparsifying Operator (SISSO) algorithm



Runhai Ouyang

R. Ouyang *et al.*, "SISSO: a compressed-sensing method for systematically identifying efficient physical models of materials properties" arxiv (2017)

SISSO overview

Step 1. Systematically construct a huge feature space

Primary features $\xrightarrow[\text{operator set}]{\textit{Feature transformation}}$ Huge feature space (billions)

$$\hat{R} = \{+, -, \times, \div, \exp, log, ^{-1}, ^{2}, ^{3}, sqrt, |-|\}$$

Step 2. Select top ranked features using Sure Independent Screening (SIS)[1]

Sure independent screening

- Select the *N* largest components of $\boldsymbol{D^T}\mathrm{y}$

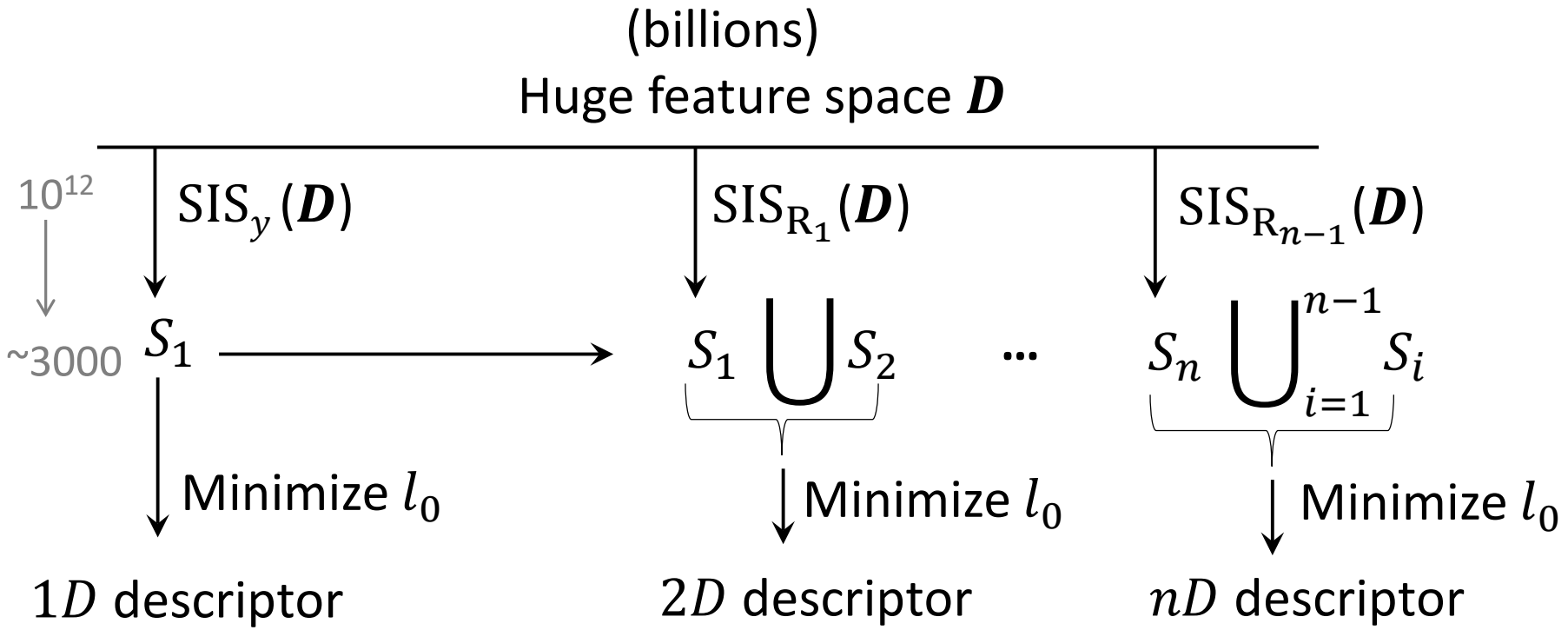
  $\boldsymbol{D}$: matrix of the features for each material

  y: target property

  Results in a subspace of *N* features that are most correlated with the property of interest

[1] J. Fan and J. Lv, J. R. Statist. Soc. B 70, 849 (2008)

Step 3. Iteratively apply sure independent screening with a sparse approximation algorithm



SIS = sure independent screening

S_i = feature subspace

y = target property

R_i = Residual of target property using the previous iterations least squares prediction

SISSO: R. Ouyang *et al.* arxiv (2017)

Also can use domain overlap for classification problems

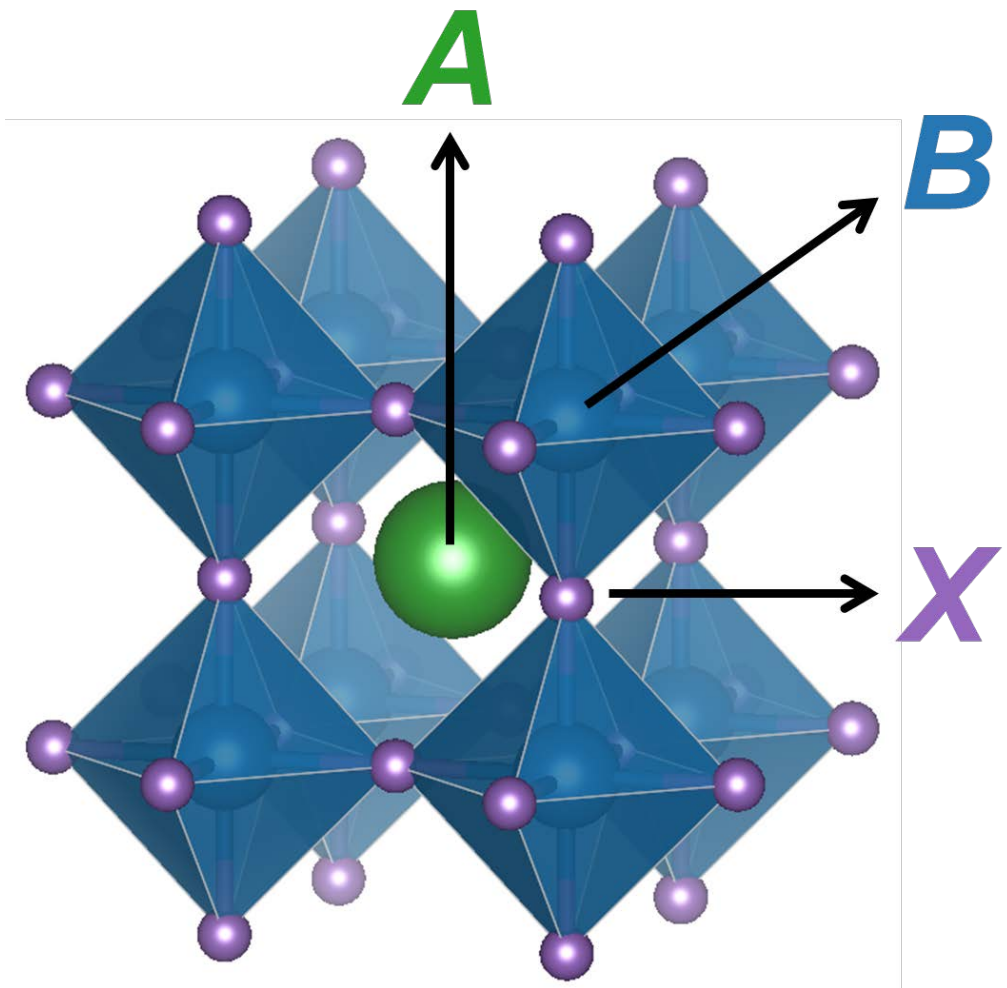
SISSO applied to perovskites to find a descriptor for their stability

Perovskites – promising functional materials

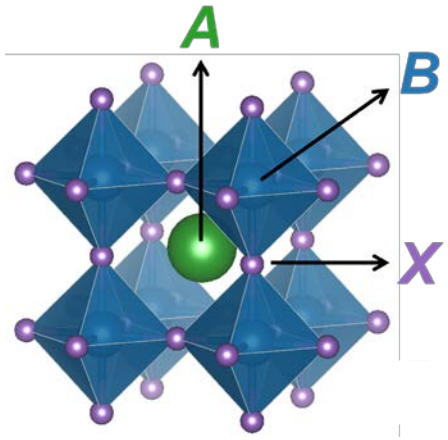
Perovskites are a class of ABX_3 materials

- *A* typically group 1, 2, or lanthanide
- *B* typically transition metal
- *X* typically chalcogen or halogen

Special Issue: Perovskites, *Science* 2017

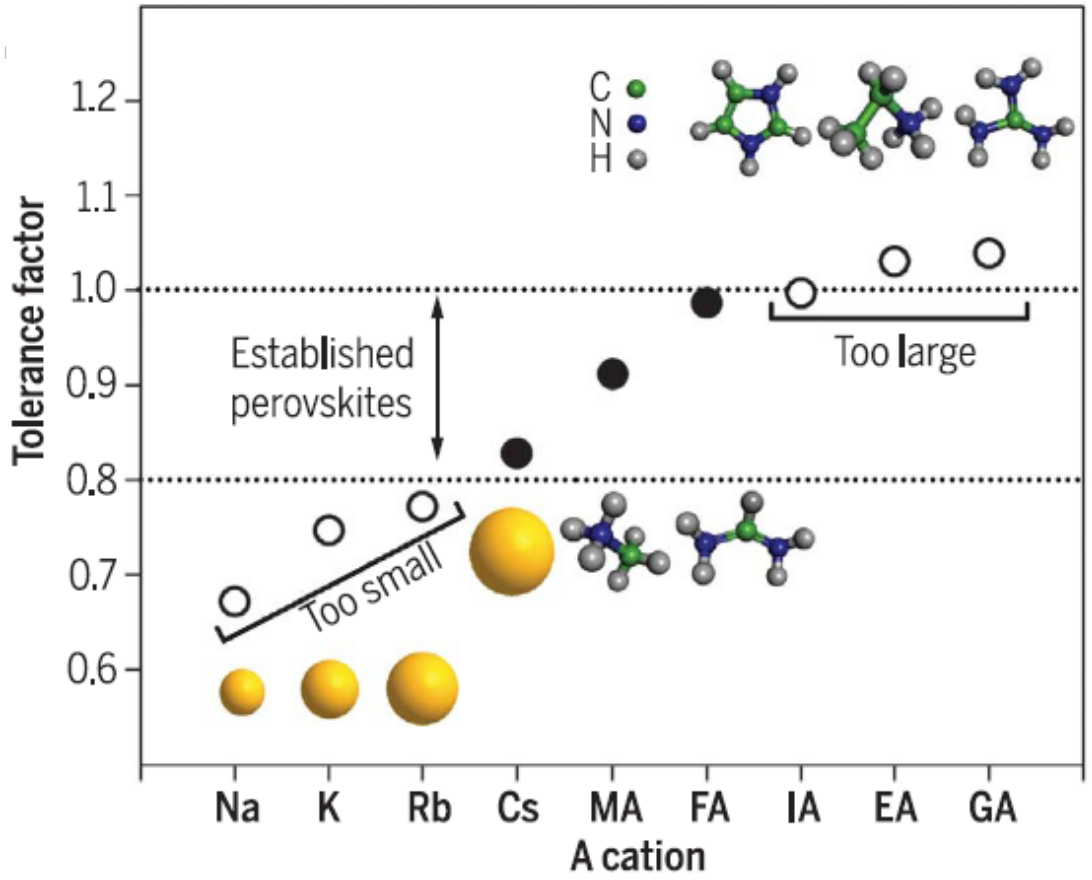


Goldschmidt's tolerance factor (t) to predict stability



$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)}$$

If cubic, $a = 2(r_B + r_X) = \sqrt{2}(r_A + r_X)$; $t = 1$



Viktor Goldschmidt (1926)

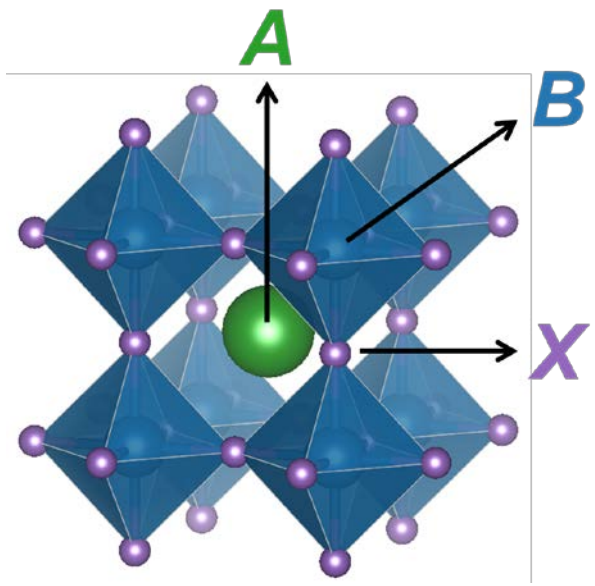
Can we find a better descriptor using SISSO?

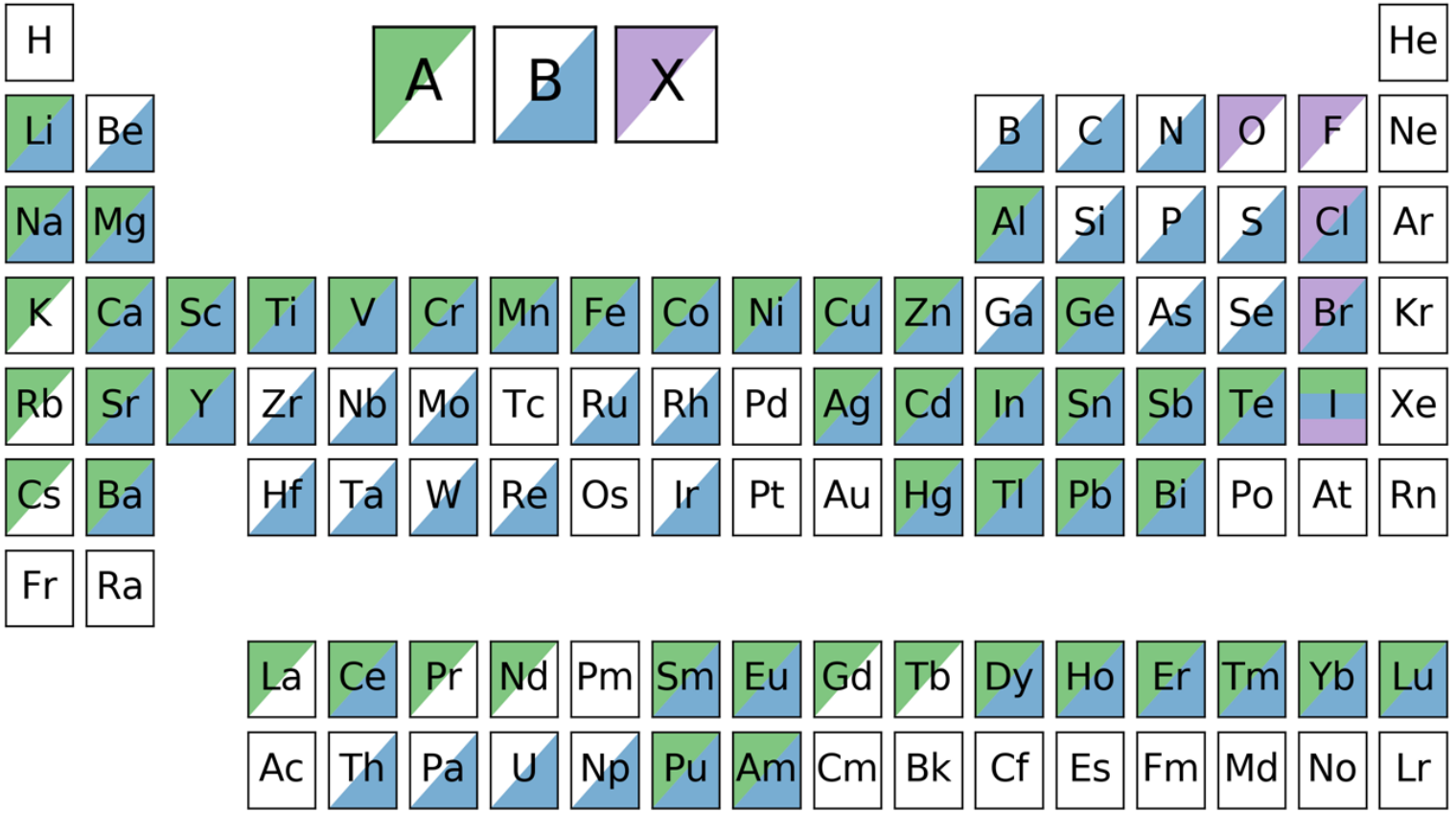
Correa-Baena *et al.*, *Science* 2017

Dataset of experimentally characterized ABX_3

576 ABX_3 with experimental XRD

- 313 perovskites and 263 nonperovskites
- 75 different cations (*A, B*)
- 5 different anions (*X*)





Experimental results compiled from:

H. Zhang, N. Li, K. Li, D. Xue, *Acta Cryst. B* 2007

C. Li, X. Lu, W. Ding, L. Feng, Y. Gao, Z. Guo, *Acta Cryst. B* 2008

W. Travis, E. Glover, H. Bronstein, D. Scanlon, R. Palgrave, *Chem. Sci.* 2016

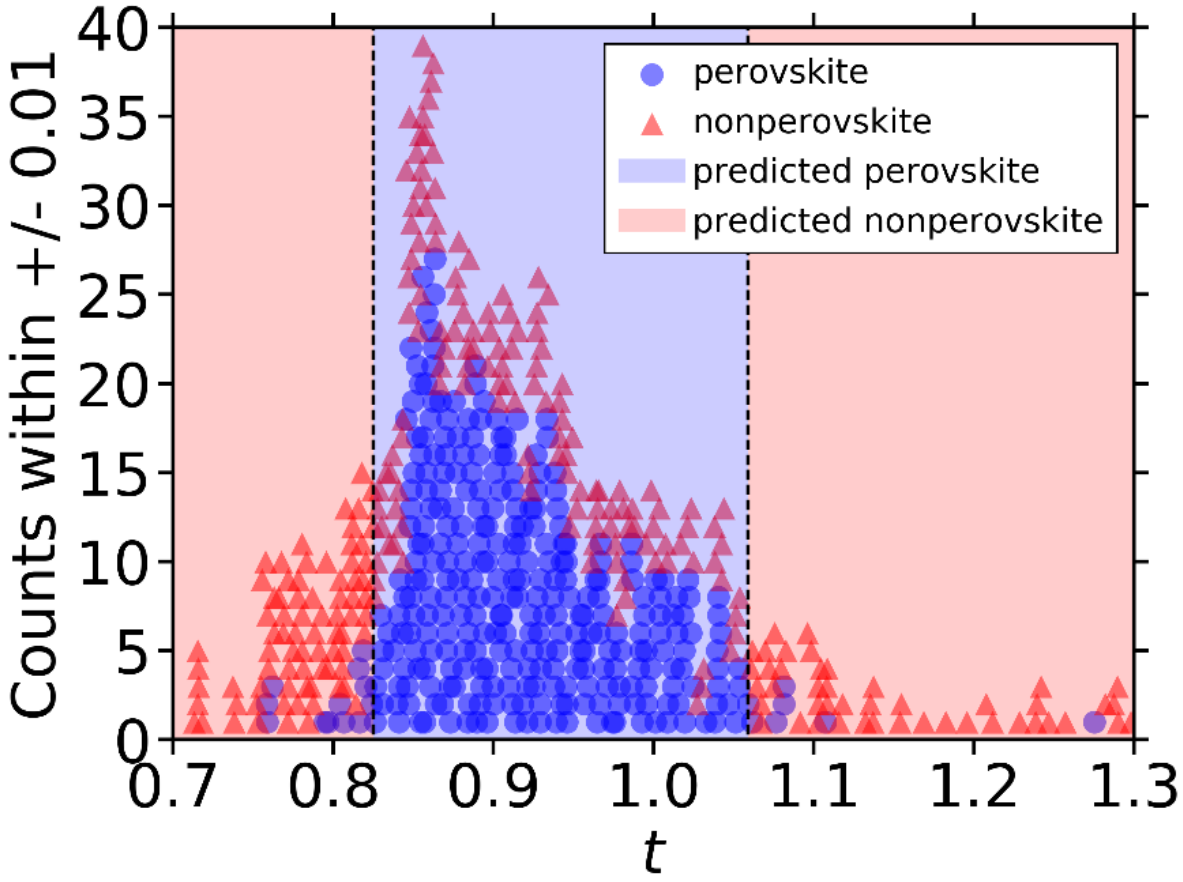
t is often insufficient, especially for halides

$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)}$$

576 ABX_3 with experimental XRD

- 313 perovskites and 263 nonperovskites
- 75 different cations (*A, B*)
- 5 different anions (*X*)

Only 74% accuracy
on experimental set



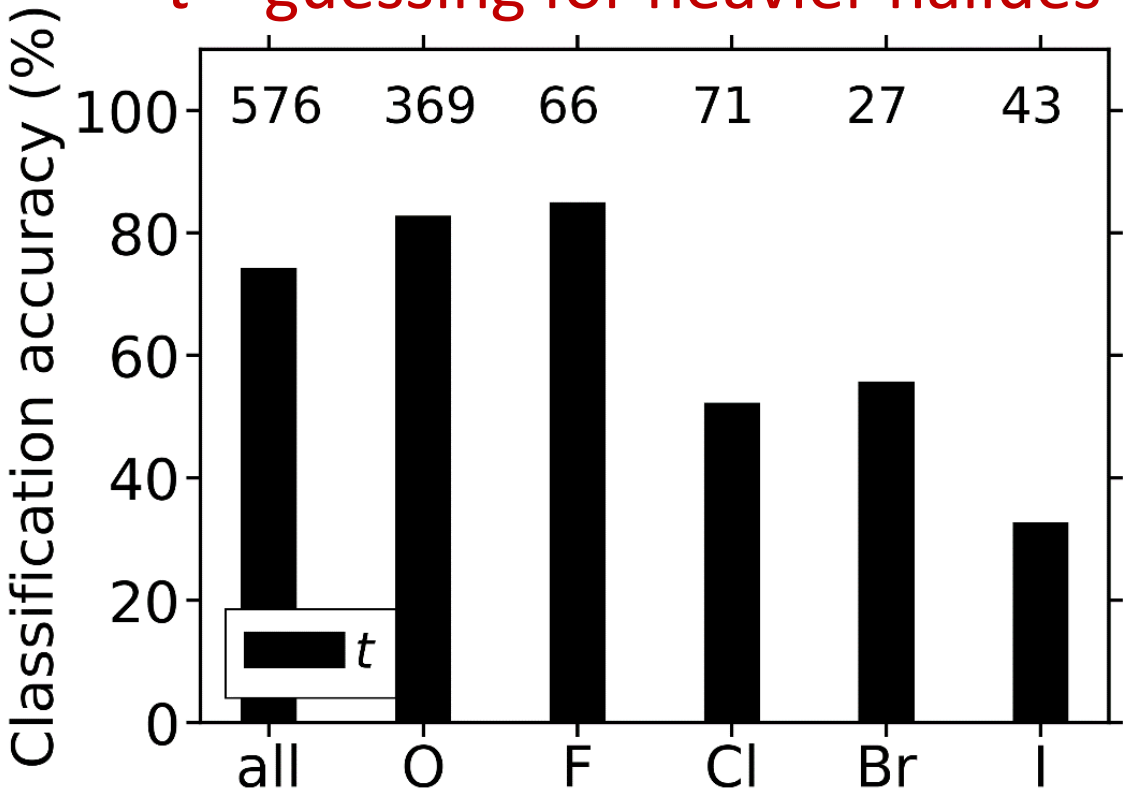
t is often insufficient, especially for halides

$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)}$$

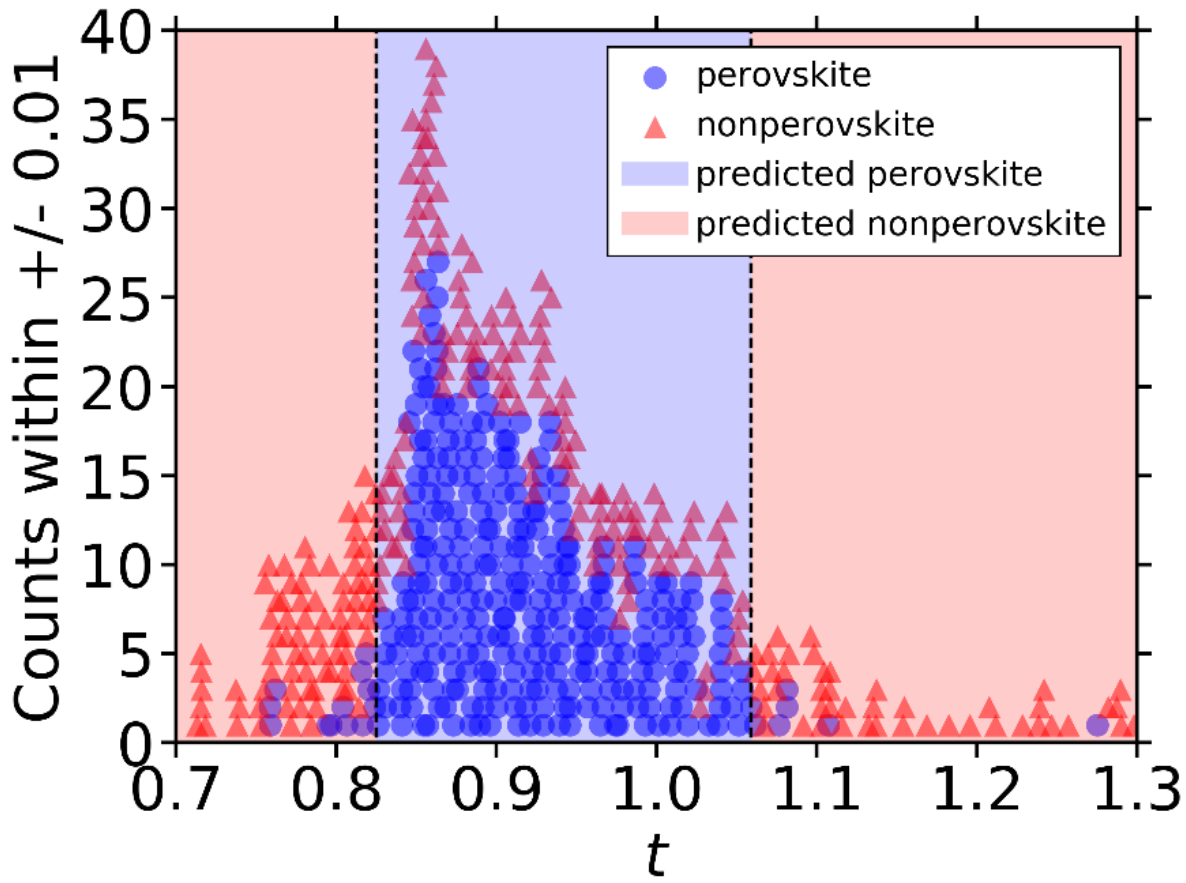
576 ABX_3 with experimental XRD

- 313 perovskites and 263 nonperovskites
- 75 different cations (*A, B*)
- 5 different anions (*X*)

t ~ guessing for heavier halides



Only 74% accuracy
on experimental set



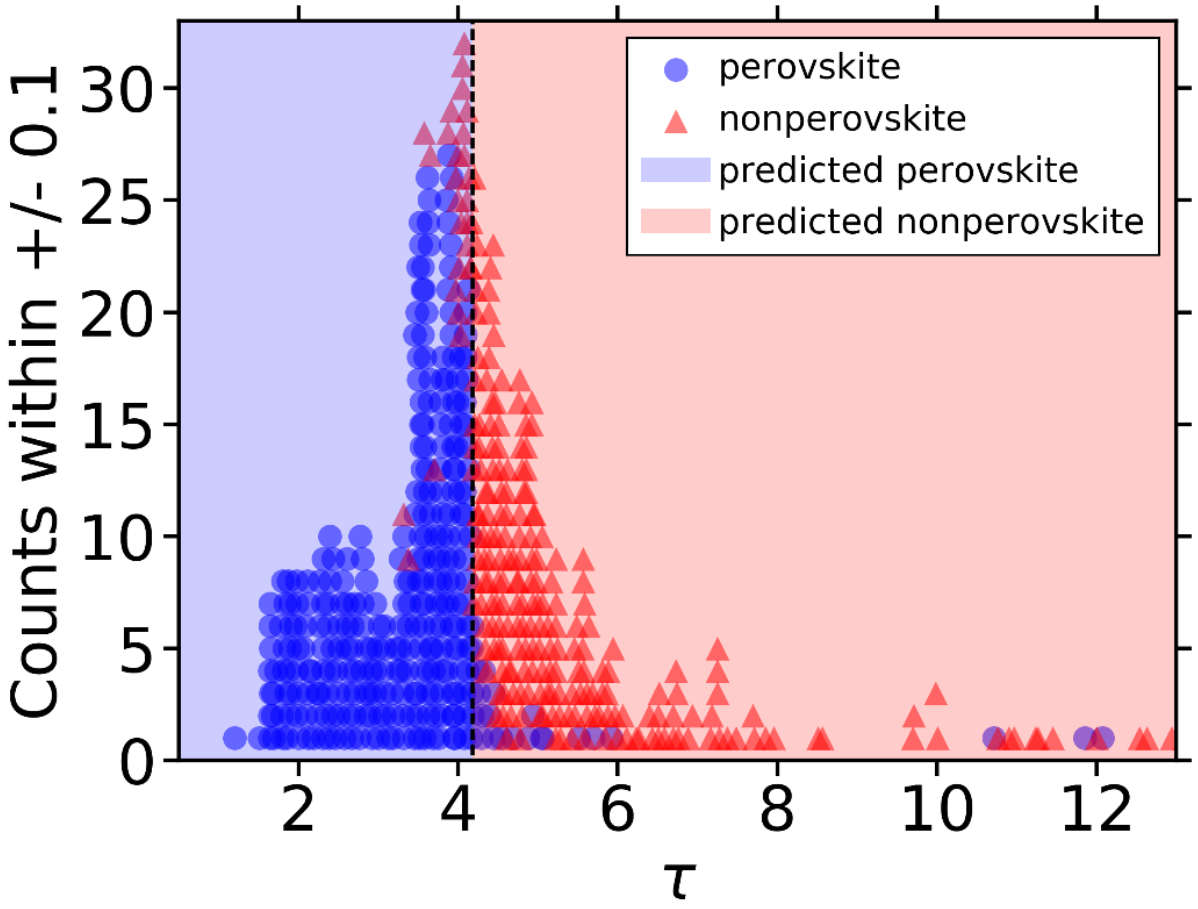
New tolerance factor discovered with SISSO (compressed sensing)

from ~3,000,000,000 potential descriptors

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

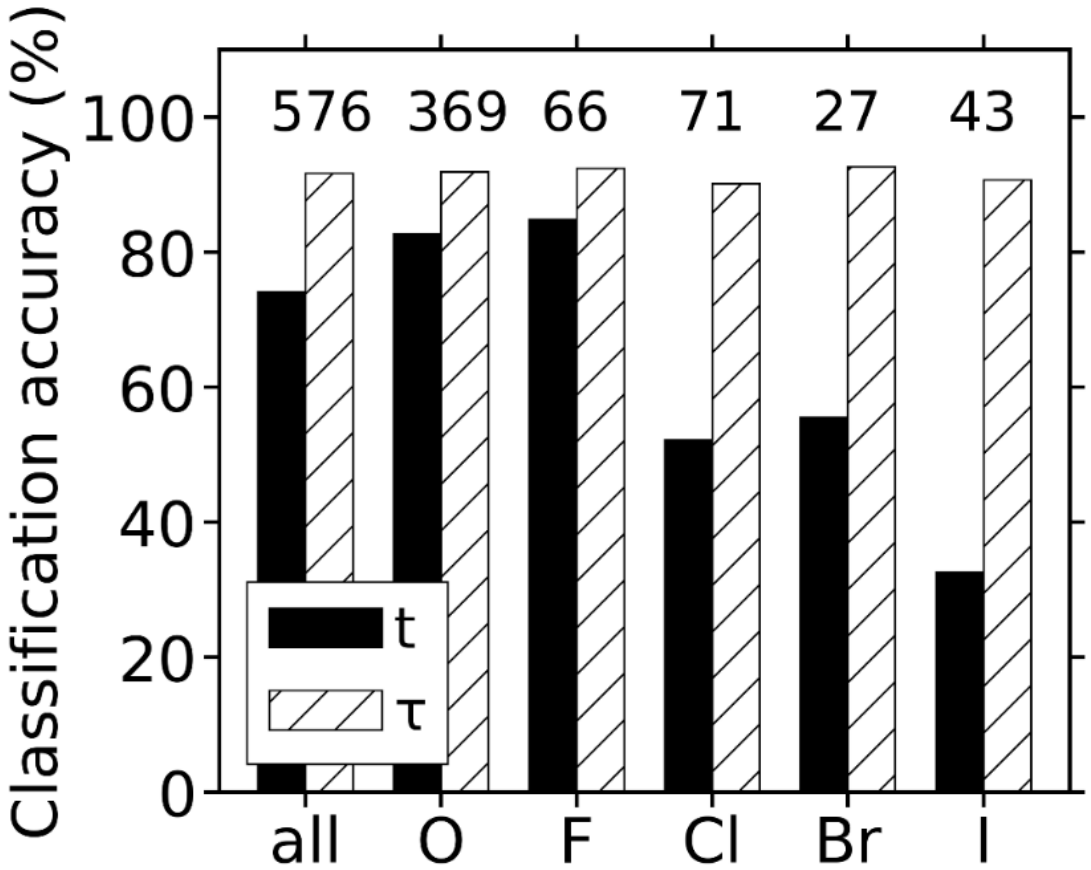
requires the same inputs as for the calculation of *t*

92% accuracy

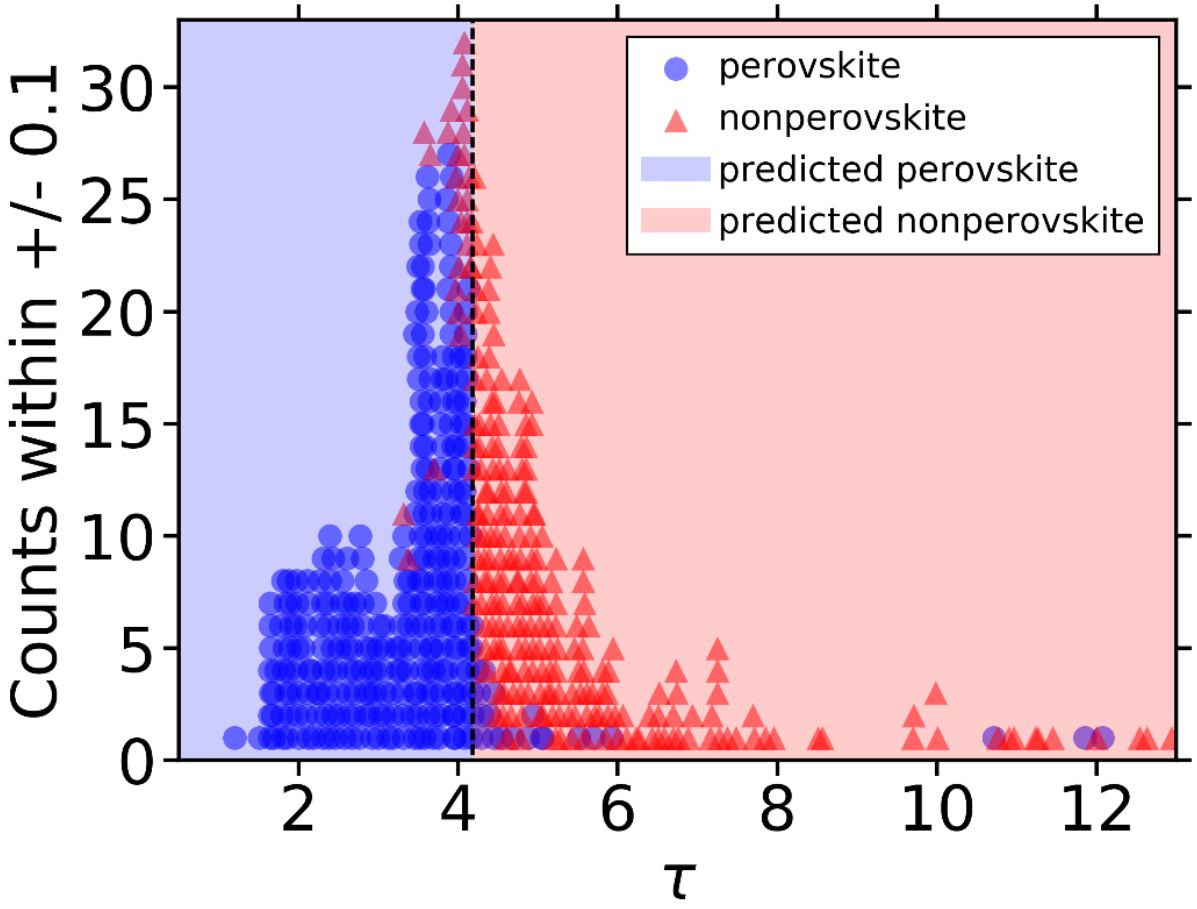


τ is general for oxides and halides

t ~ uniform performance



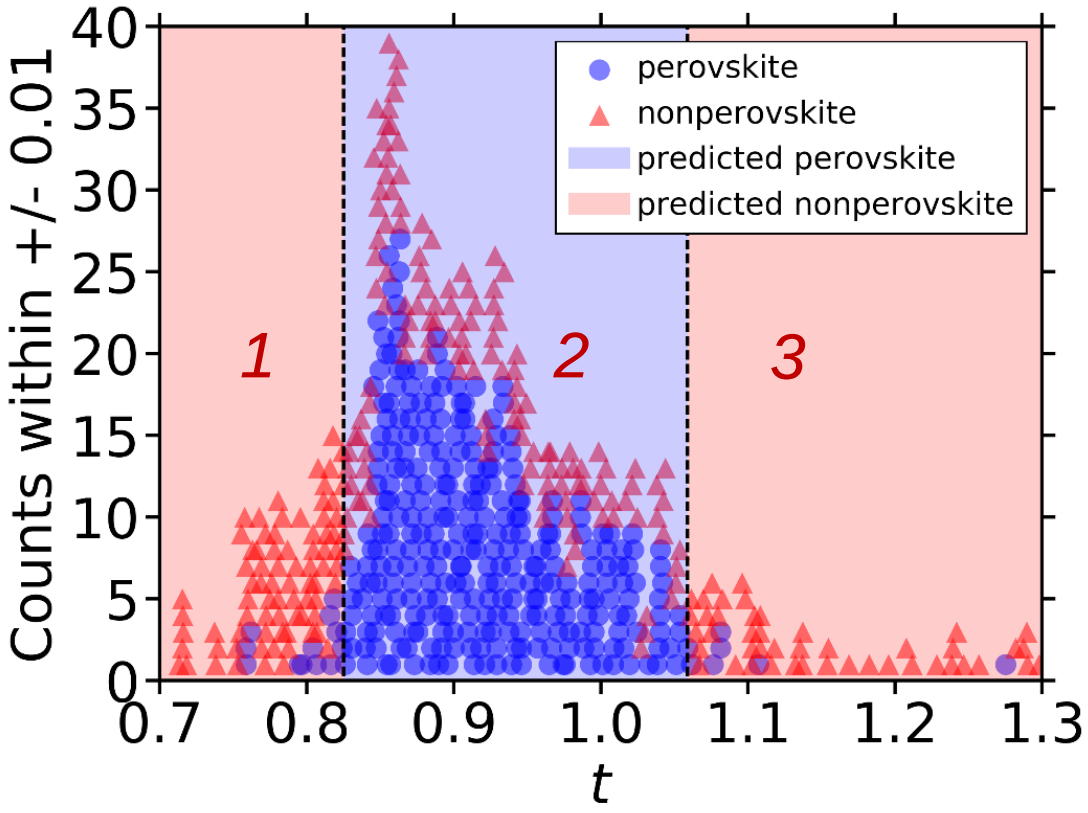
92% accuracy



Comparing the forms of t and τ

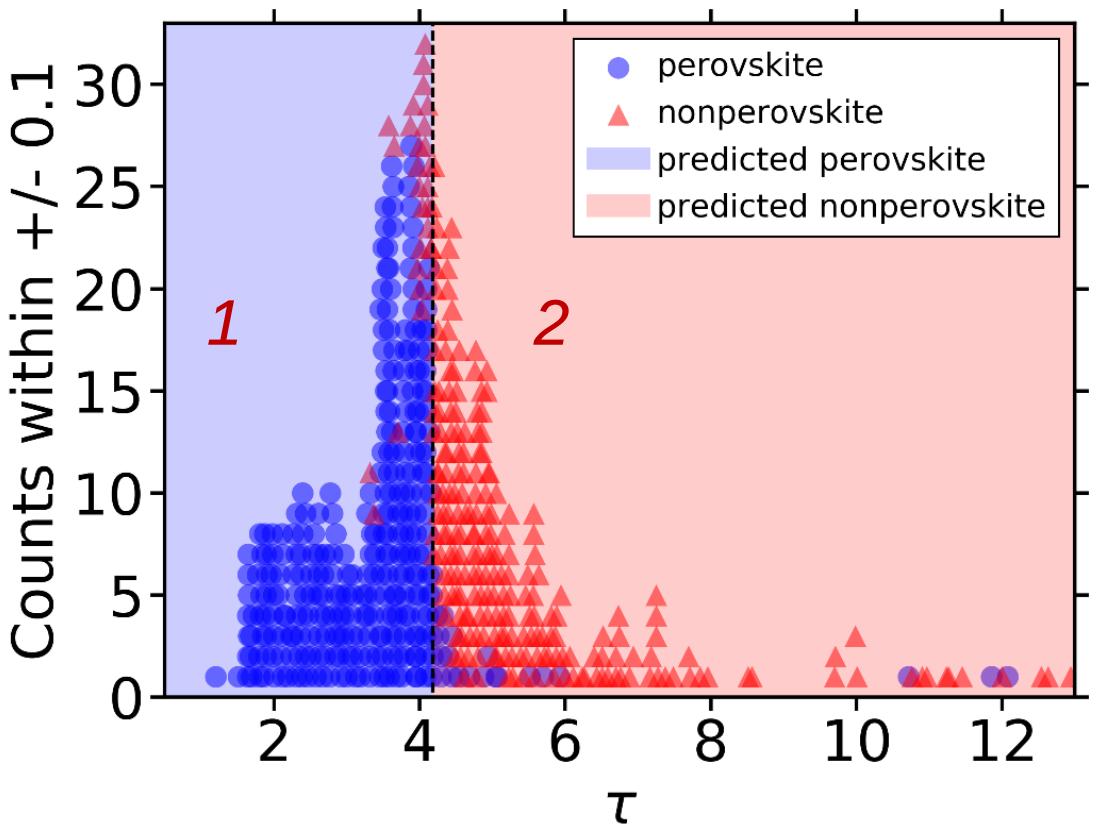
$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)}$$

$$\tau = \frac{r_X}{r_B} - n_A\left(n_A - \frac{r_A/r_B}{ln\, r_A/r_B}\right)$$



$0.825 < t < 1.059 \rightarrow$ perovskite

74% accuracy



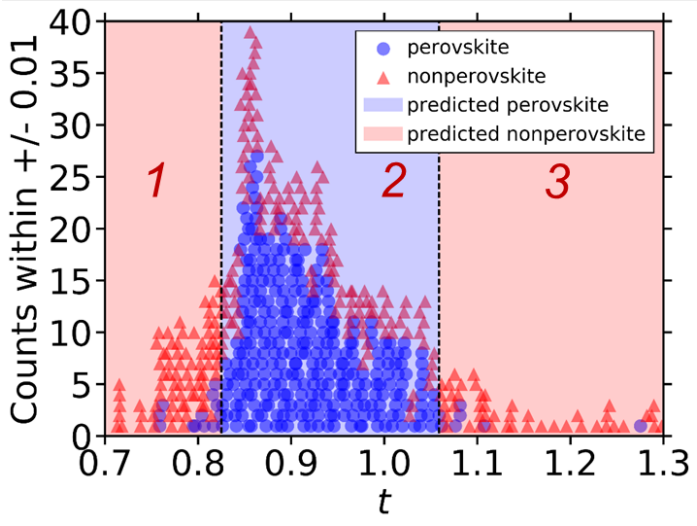
$\tau < 4.18 \rightarrow$ perovskite

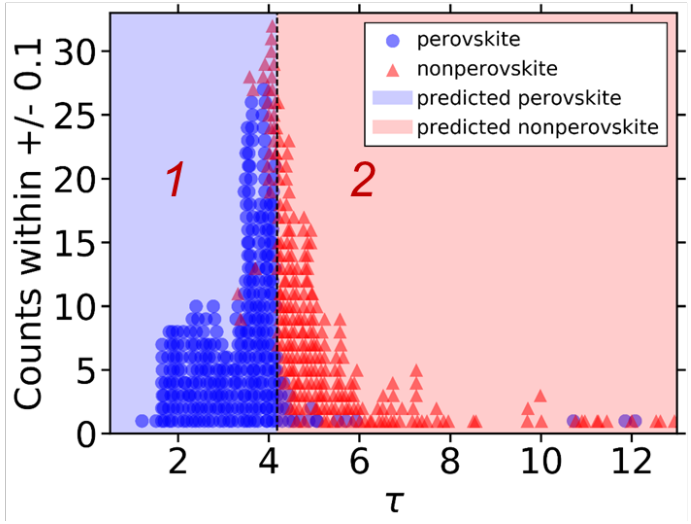
92% accuracy

Monotonic perovskite probabilities – $\wp(\tau)$

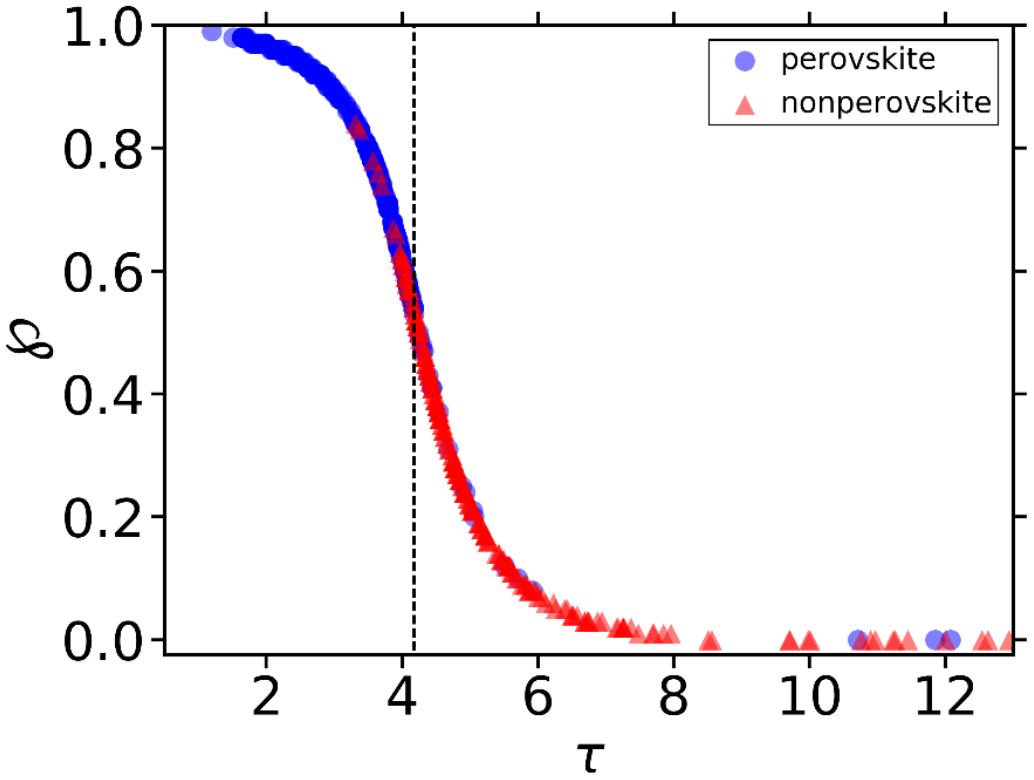
$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)}$$

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





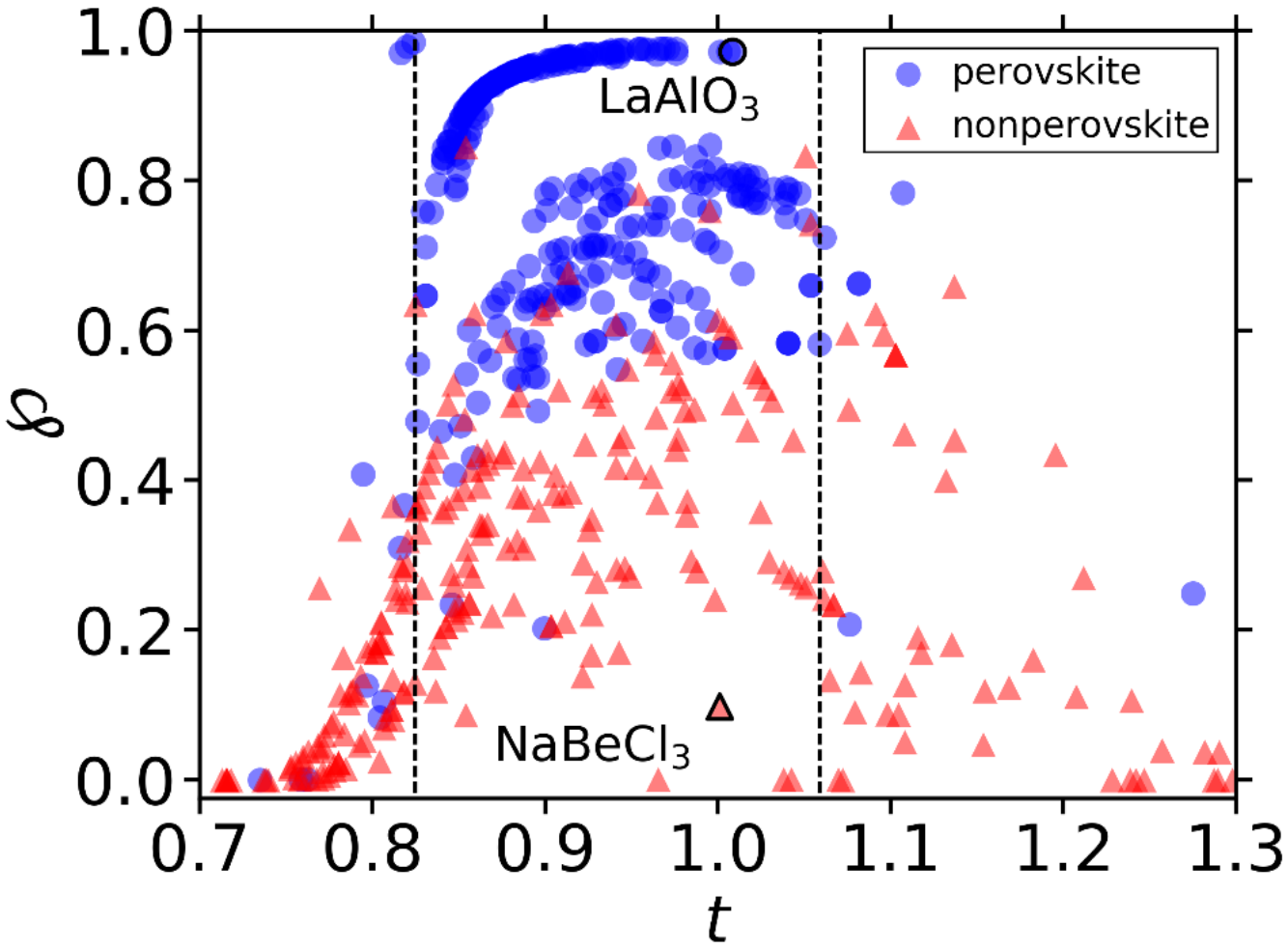
Decision tree outputs only (-1, 1)



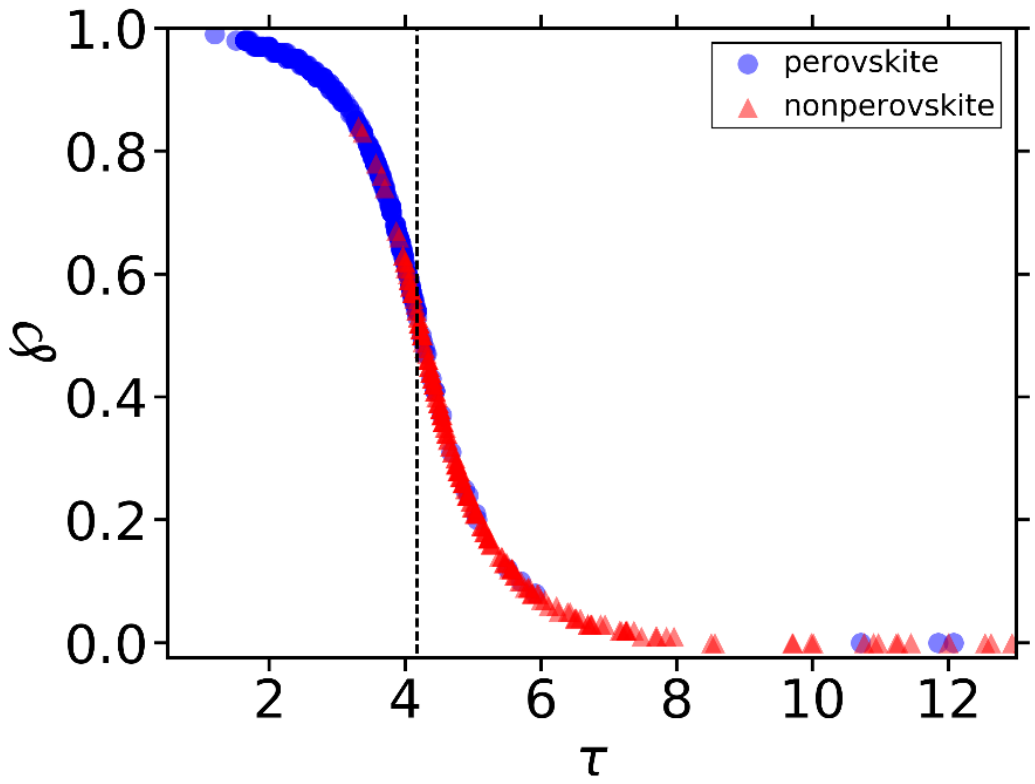
Mapping decision tree outputs to logistic regression[1] yields $\wp(\tau)$

[1]J. Platt, *Advances In large margin classifiers*, 1999

Monotonic perovskite probabilities – $\wp(\tau)$



Decision tree outputs only (-1, 1)

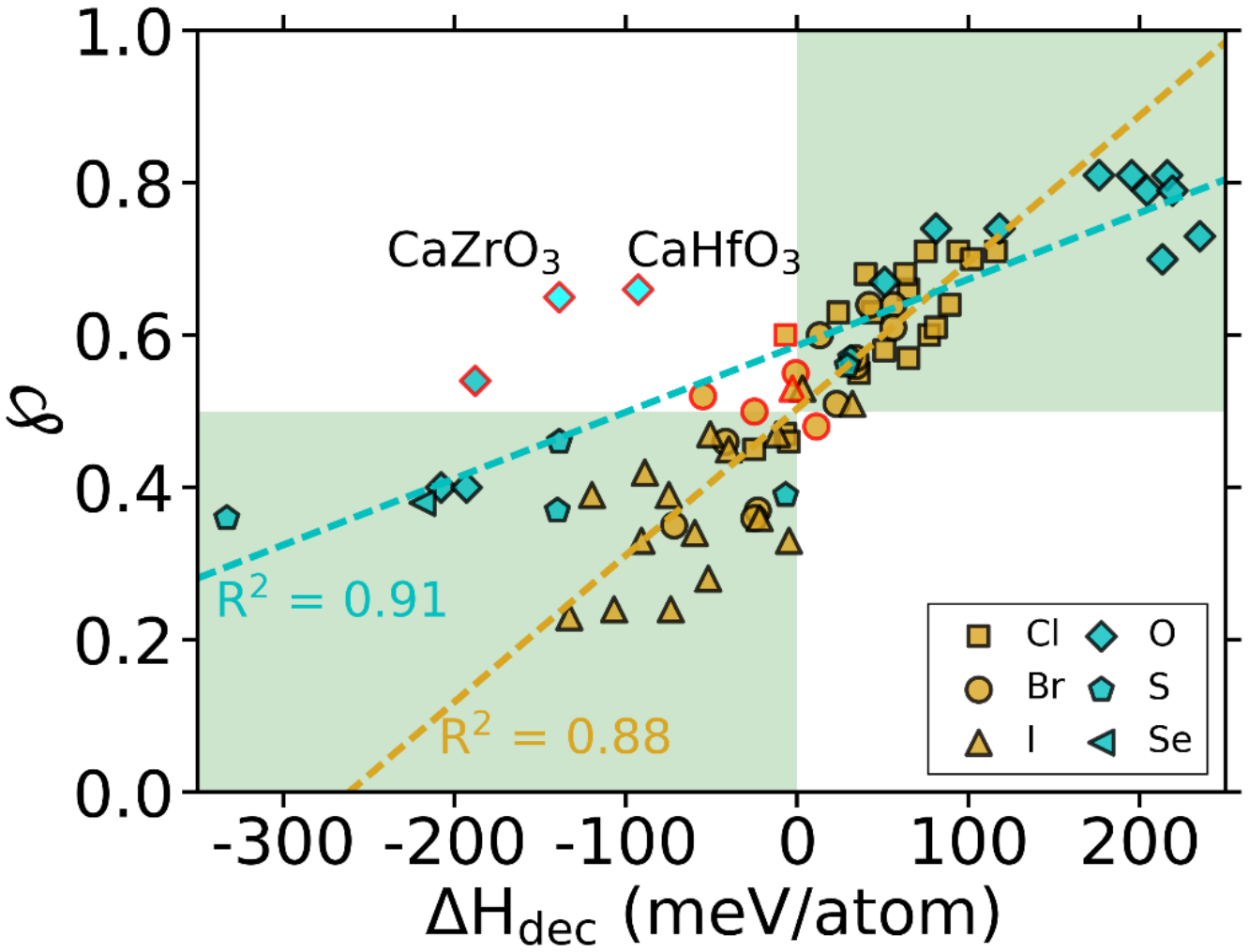


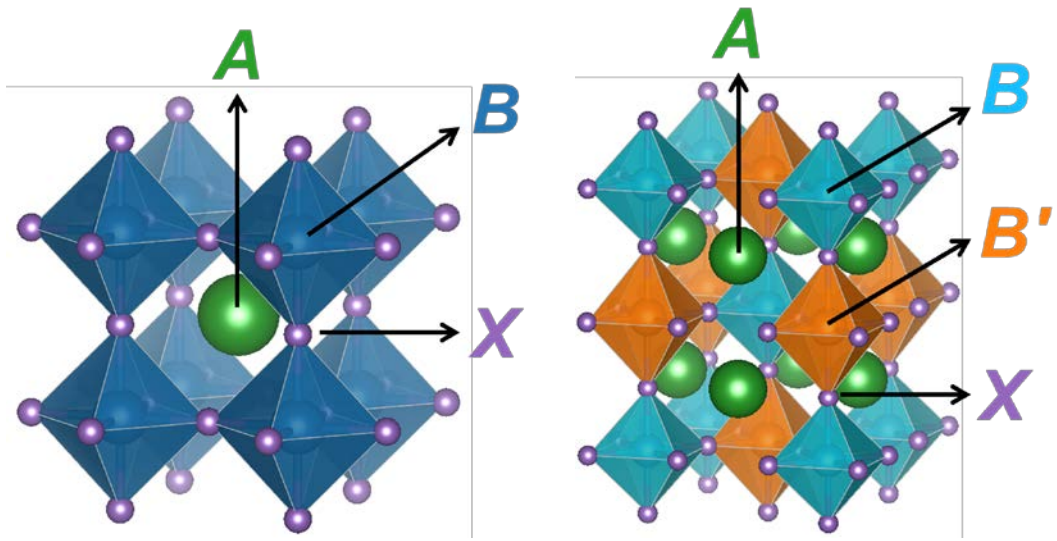
Mapping decision tree outputs to logistic regression[1] yields $\wp(\tau)$

[1]J. Platt, *Advances In large margin classifiers*, 1999

$\wp(\tau)$ compares well with DFT-GGA ΔH_{dec}

$\Delta H_{dec} > 0 \rightarrow$ stable in cubic structure





88% agreement

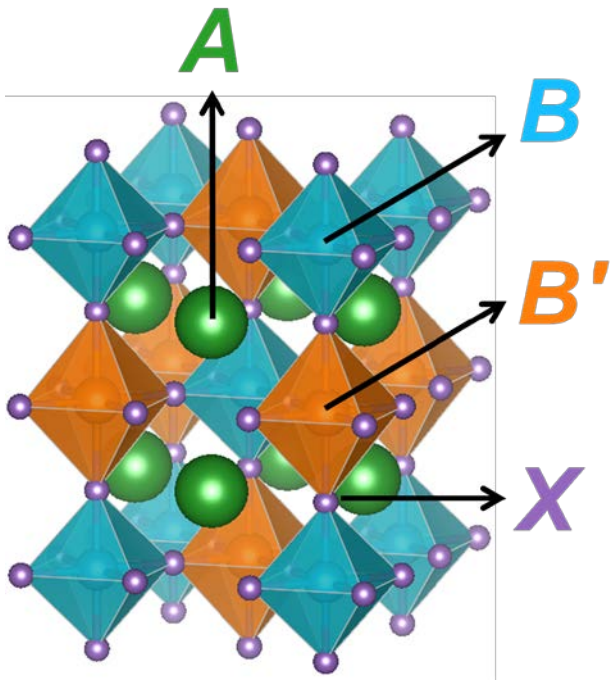
$\wp$ correlates with ΔH_{dec}

τ *can* be more powerful ($CaZrO_3$, $CaHfO_3$)

Decomposition enthalpies from:
X.-G. Zhao, D. Yang, Y. Sun, T. Li, L. Zhang, L. Yu, A. Zunger, *JACS* 2017
Q. Sun, W.-J. Yin, *JACS* 2017

Double perovskites for emerging solar absorbers

$A_2BB'X_6$



9 recently synthesized double perovskite halides – *all* predicted to be perovskite by τ

J|A|C|S JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
Communication
pubs.acs.org/JACS

2016

A Bismuth-Halide Double Perovskite with Long Carrier Recombination Lifetime for Photovoltaic Applications

Adam H. Slavney,[†] Te Hu,[§] Aaron M. Lindenberg,[§] and Hemamala I. Karunadasa*[,†]

THE JOURNAL OF PHYSICAL CHEMISTRY Letters
Letter
pubs.acs.org/JPCL

2016

Lead-Free Halide Double Perovskites via Heterovalent Substitution of Noble Metals

George Volonakis,[†] Marina R. Filip,[†] Amir Abbas Haghighirad,[‡] Nobuya Sakai,[‡] Bernard Wenger,[‡] Henry J. Snaith,*[,‡] and Feliciano Giustino*[,†]

J|A|C|S JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
Article
pubs.acs.org/JACS

2017

Cu–In Halide Perovskite Solar Absorbers

Xin-Gang Zhao,[†] Dongwen Yang,[†] Yuanhui Sun,[†] Tianshu Li,[†] Lijun Zhang,*[,†] Liping Yu,[‡] and Alex Zunger[§]

THE JOURNAL OF PHYSICAL CHEMISTRY Letters
Letter
Cite This: J. Phys. Chem. Lett. 2017, 8, 5015-5020
pubs.acs.org/JPCL

2017

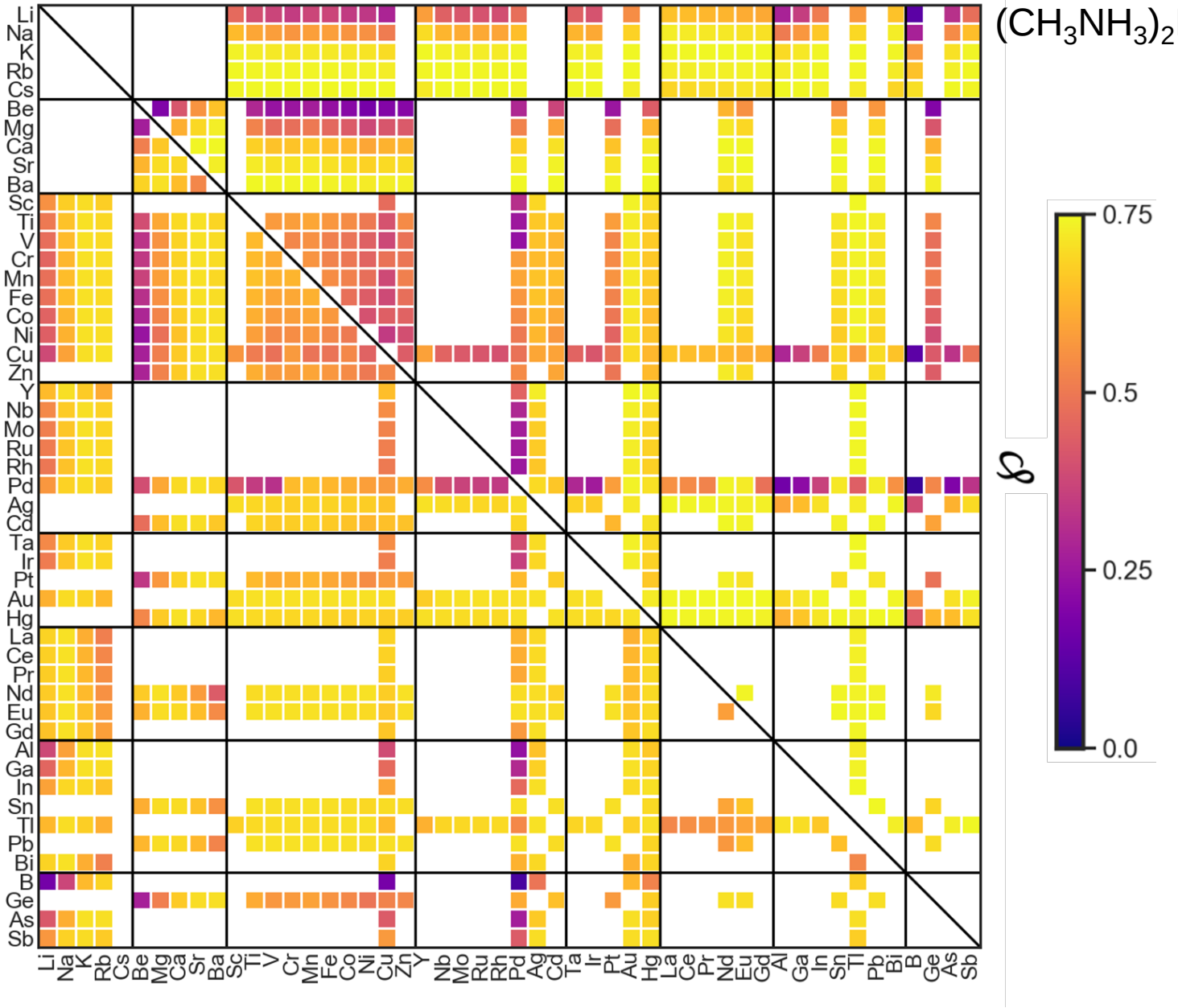
Synthesis and Characterization of the Rare-Earth Hybrid Double Perovskites: $(CH_3NH_3)_2KGdCl_6$ and $(CH_3NH_3)_2KYCl_6$

Zeyu Deng,[†] Fengxia Wei,[†,‡] Federico Brivio,[†] Yue Wu,[†] Shijing Sun,[†] Paul D. Bristowe,*[,†] and Anthony K. Cheetham*[,†]

... and more every month

τ applied to 259,296 $A_2BB'X_6$ compounds

Upper triangle – $(CH_3NH_3)_2BB'Br_6$

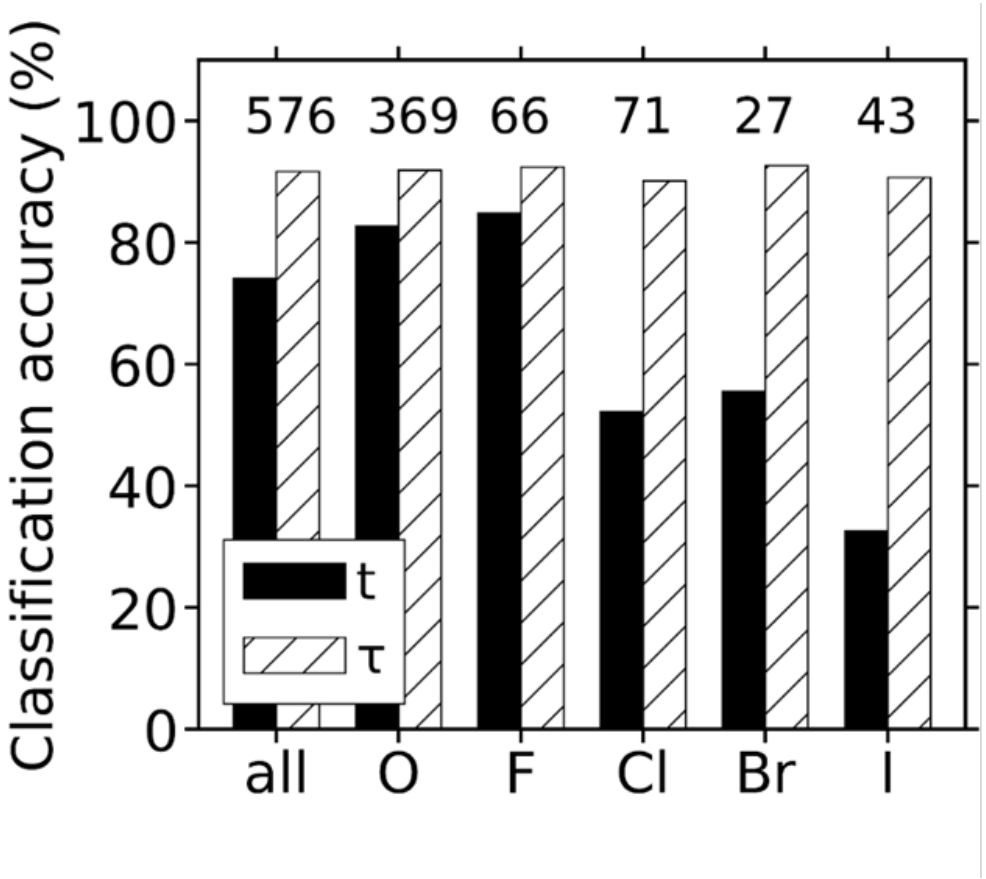


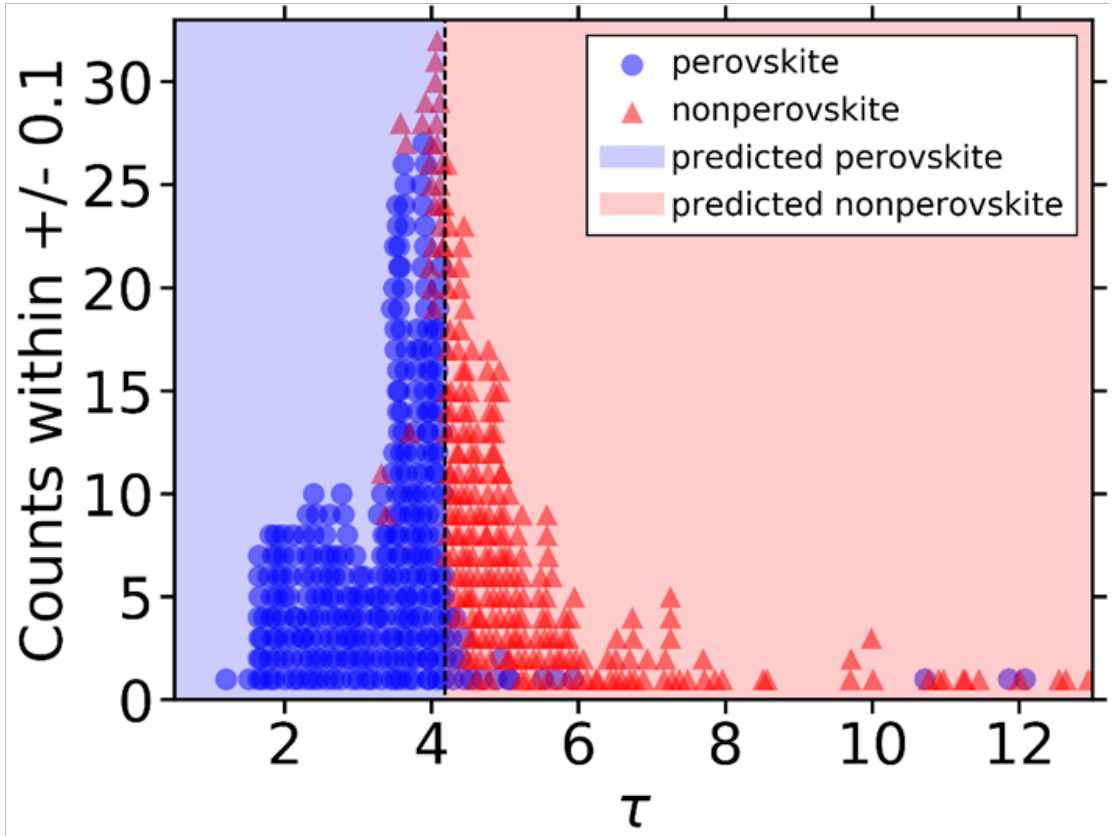
Lower triangle – $Cs_2BB'Cl_6$

New tolerance factor for perovskite stability using SISSO

- SISSO to find new descriptor, τ, which improves upon Goldschmidt's

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



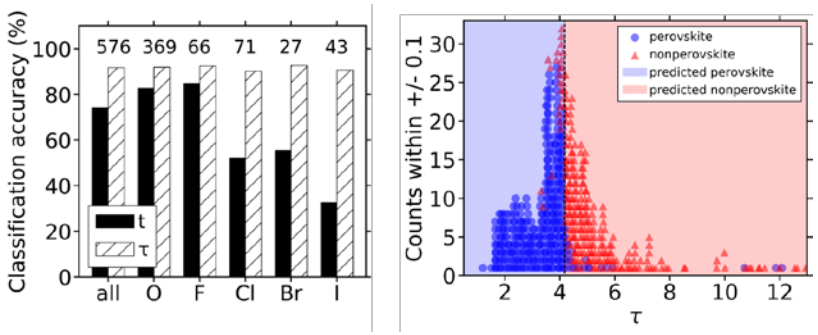


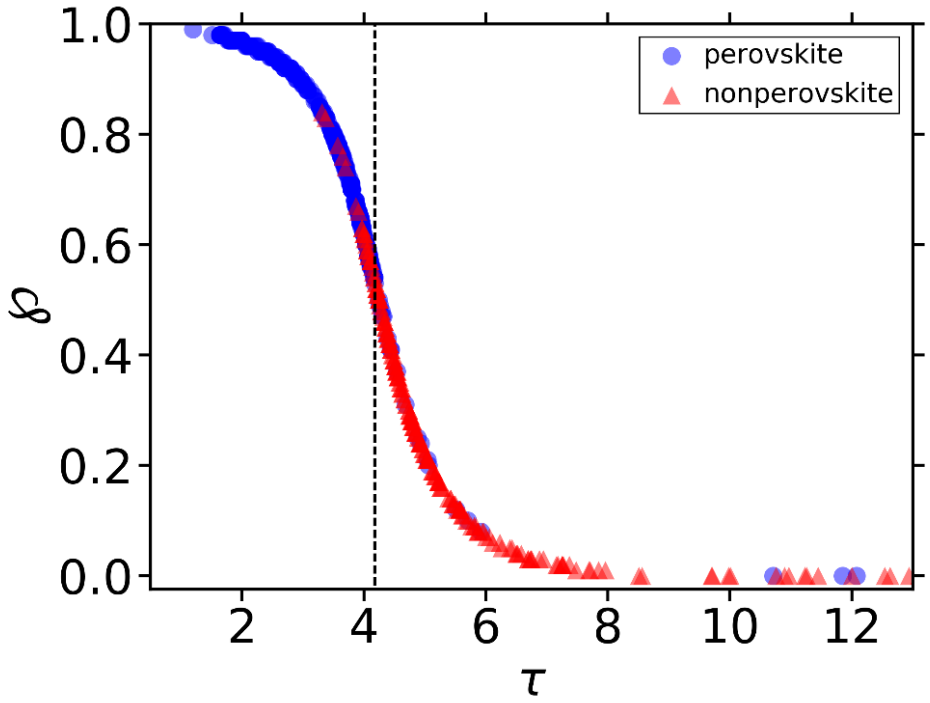
C. Bartel *et al.*, *Under review* (https://arxiv.org/abs/1801.07700)

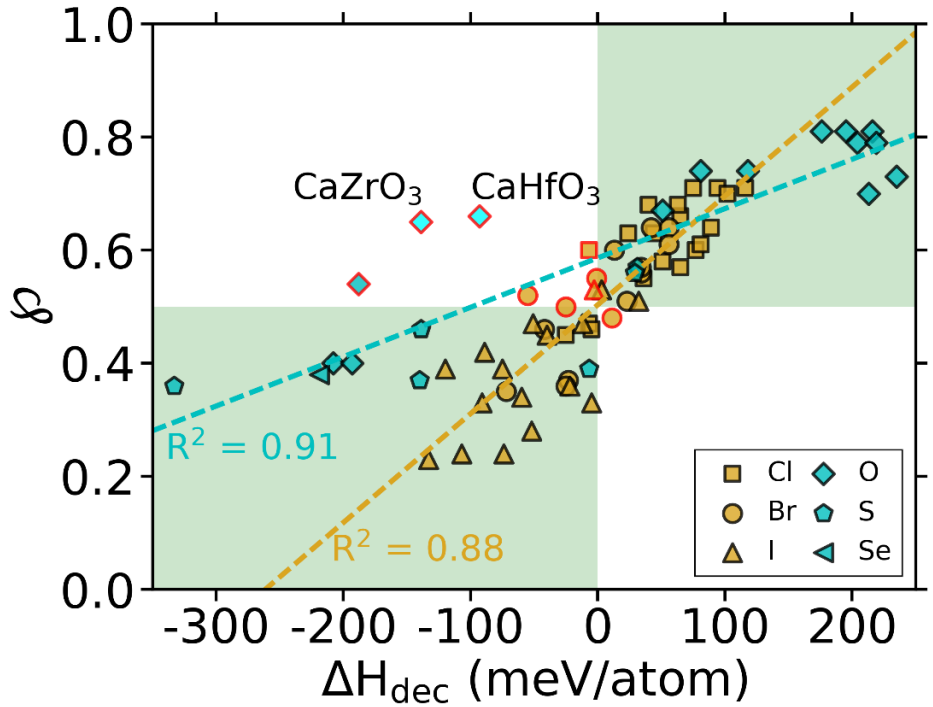
github.com/CJBartel/perovskite-stability

New tolerance factor for perovskite stability using SISSO

- SISSO to find new descriptor, τ, which improves upon Goldschmidt's
- τ yields meaningful probabilities





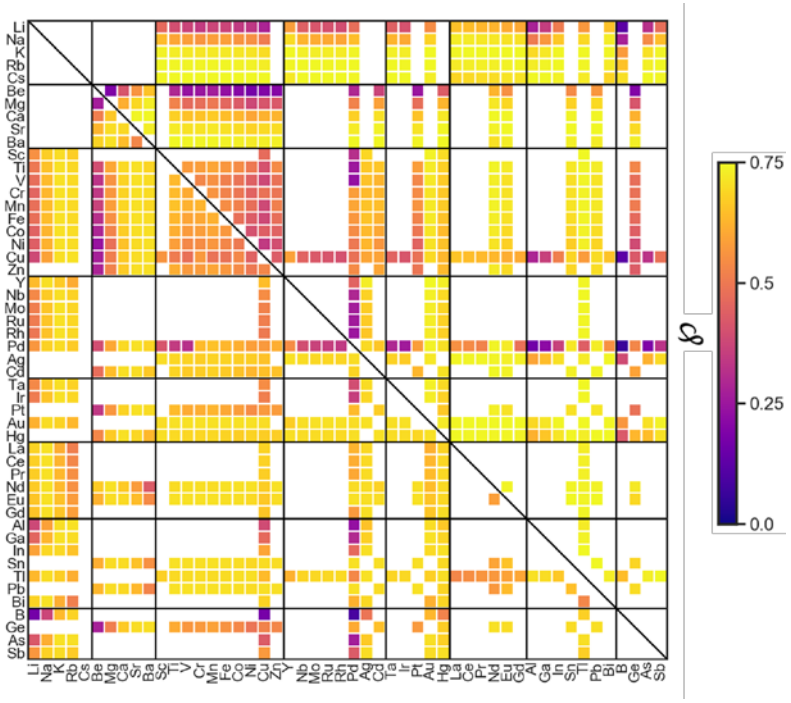


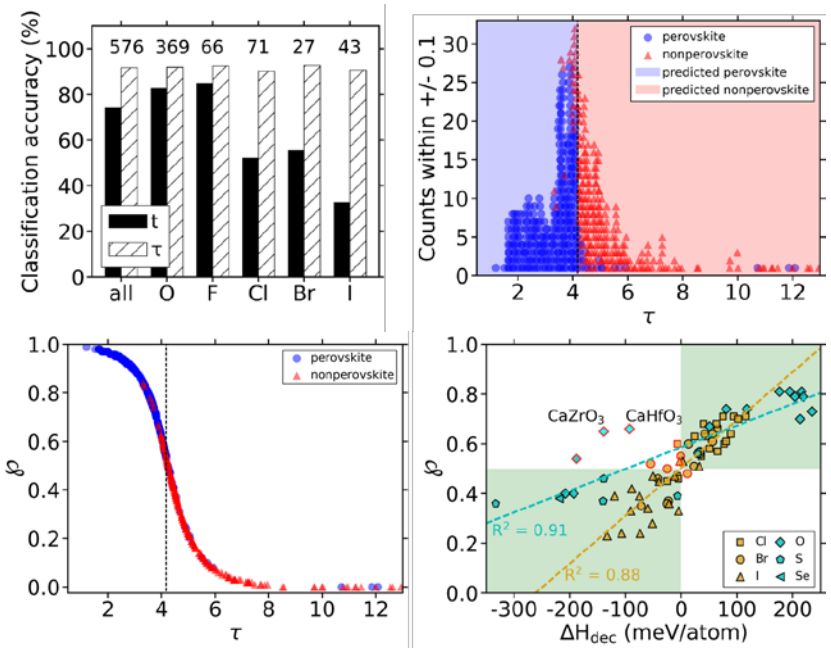
C. Bartel *et al.*, *Under review* (https://arxiv.org/abs/1801.07700)

github.com/CJBartel/perovskite-stability

New tolerance factor for perovskite stability using SISSO

- SISSO to find new descriptor, τ, which improves upon Goldschmidt's
- τ yields meaningful probabilities
- Stability elucidated as $\wp(A, B, X)$



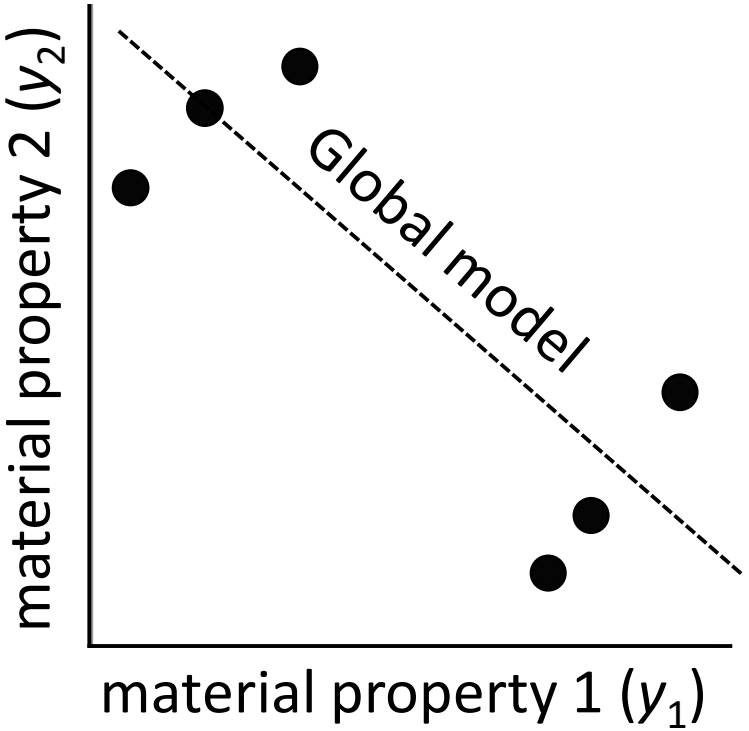


SISSO should be applicable to catalysis problems

C. Bartel *et al.*, *Under review* (https://arxiv.org/abs/1801.07700)

Part 2. *Subgroup discovery* to find local patterns and their descriptions

Typically one focuses on creating a global prediction model for some property of interest (e.g., SISSO, Kernel Ridge Regression, Neural Networks)

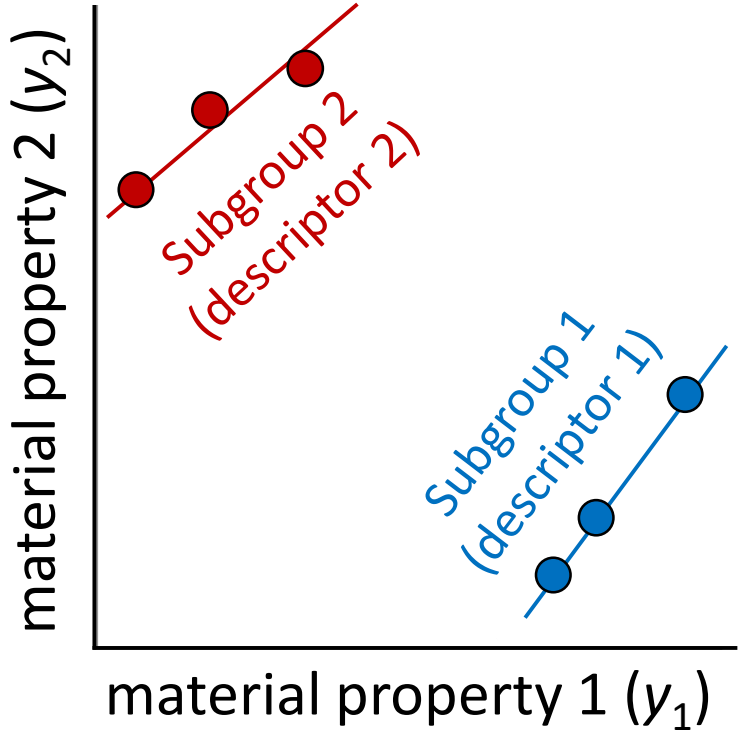


Underlying mechanisms can change across materials

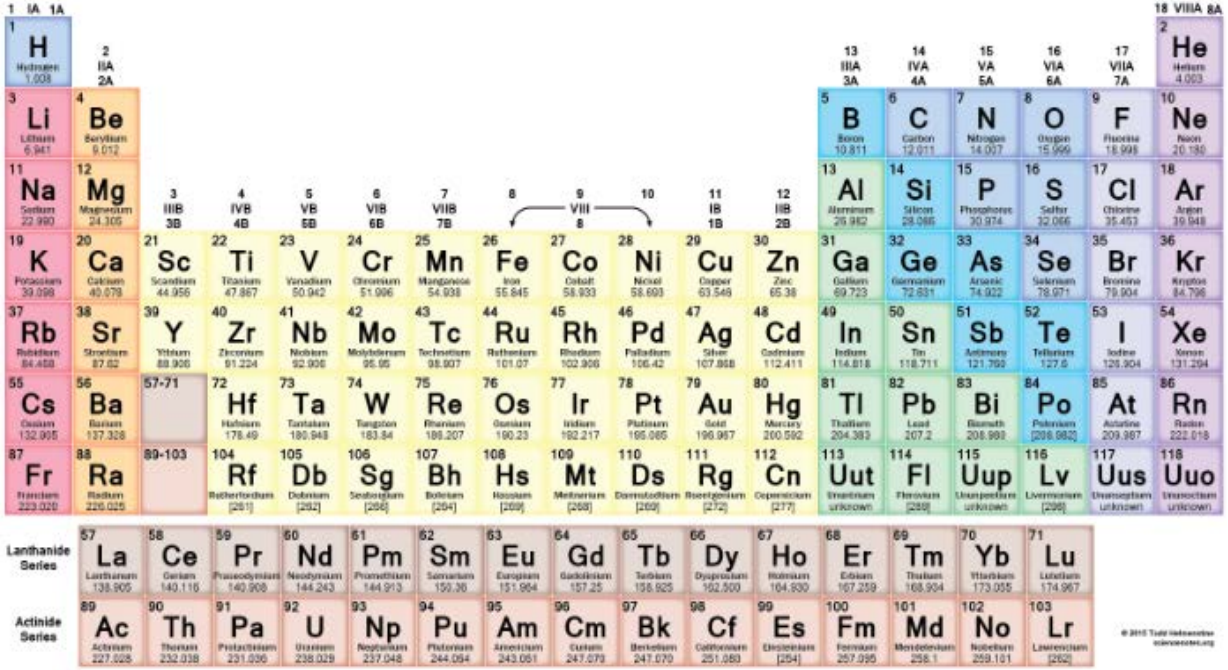
Relations between subsets of data may be important

Subgroup discovery:

find meaningful *local descriptors*
of a target property in materials-science data



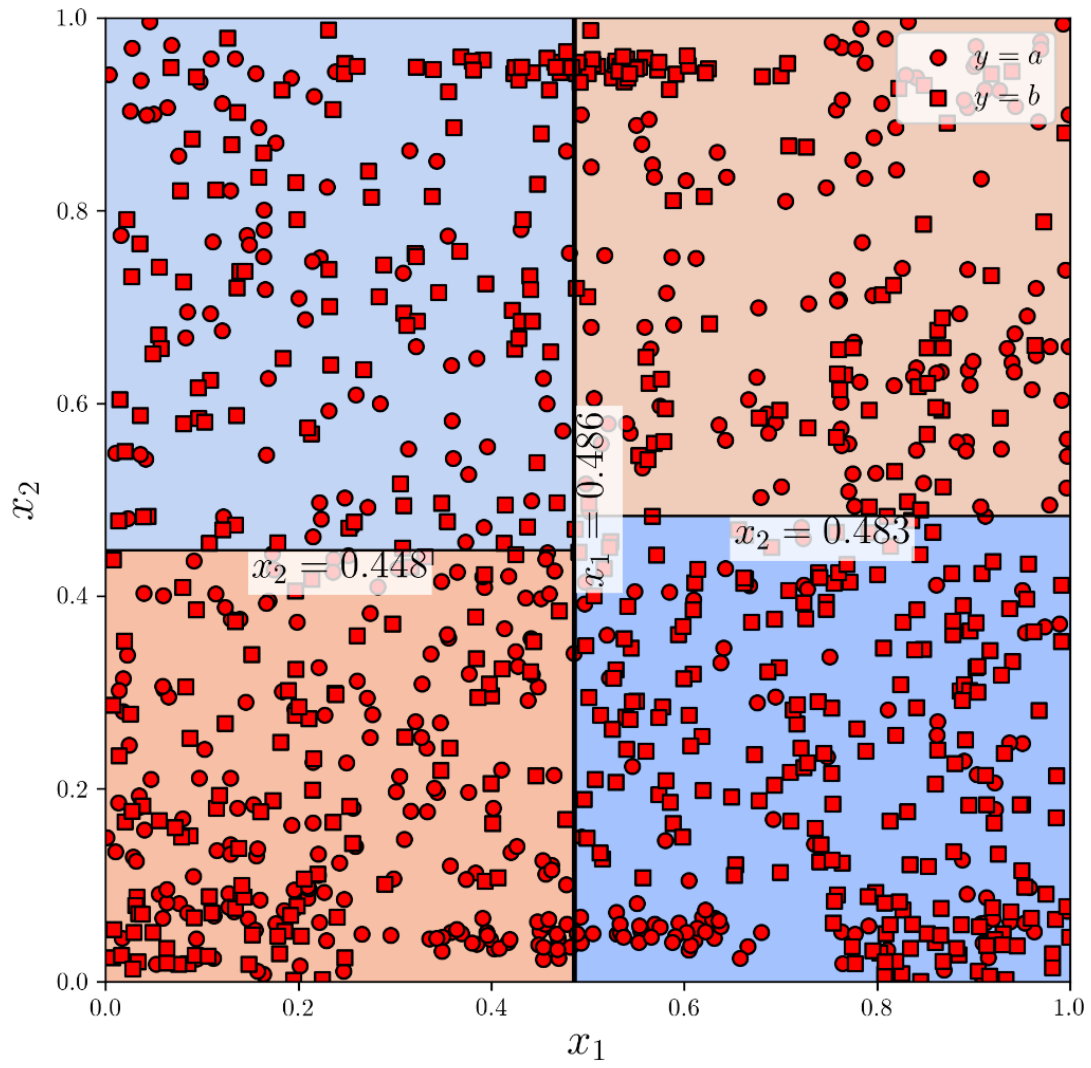
The periodic table has subgroups



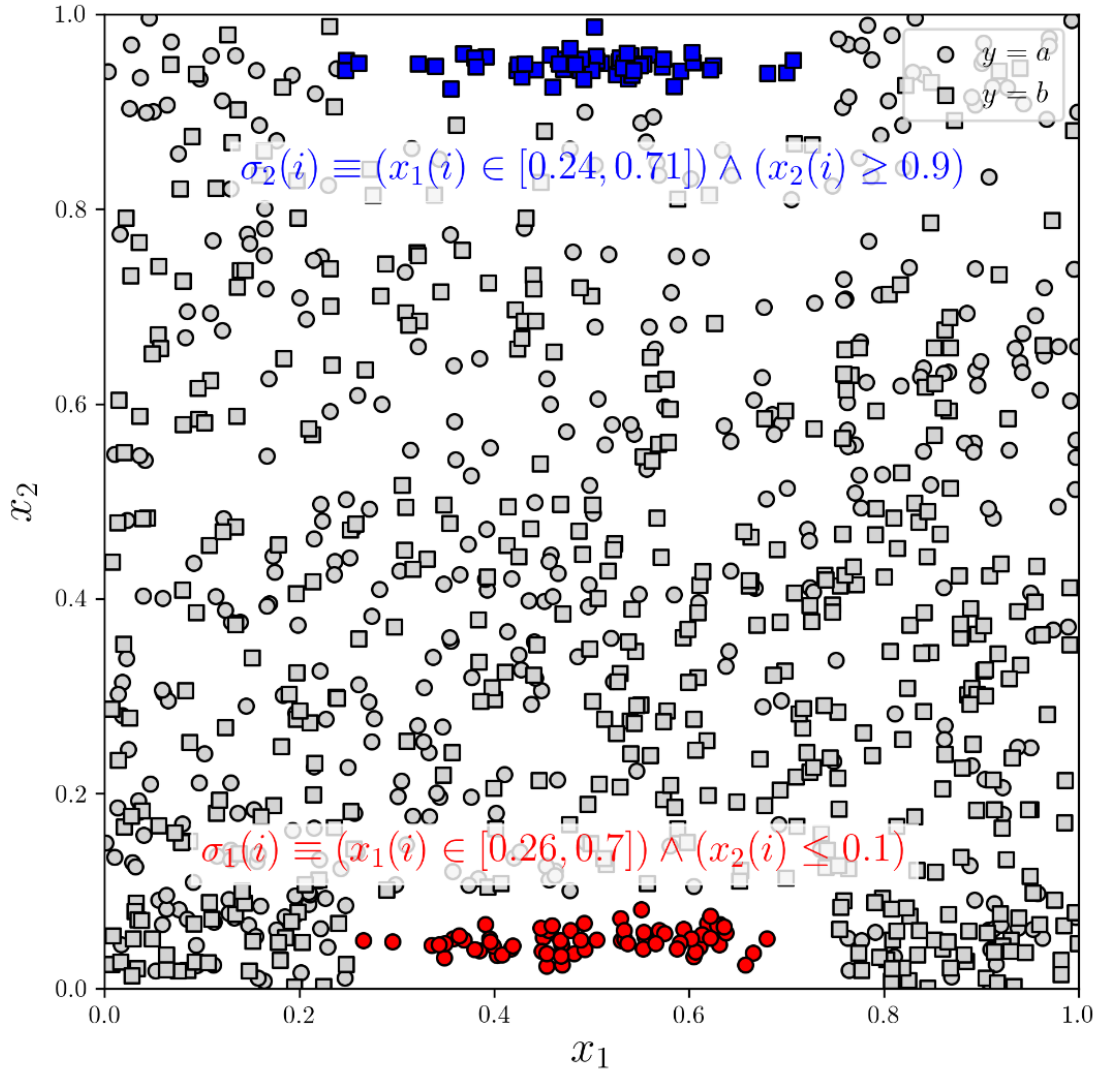
Review: M. Atzmueller, *WIREs Data Min. Knowl. Disc. 5* (2015)
B. R. Goldsmith, M. Boley, J. Vreeken, M. Scheffler, L. M. Ghiringhelli, *New J. Phys.* 19, (2017)
M. Boley, B. R. Goldsmith, L. M. Ghiringhelli, J. Vreeken, *Data Min. Knowl. Disc.* 1391, (2017)

Subgroup discovery focuses on local observations

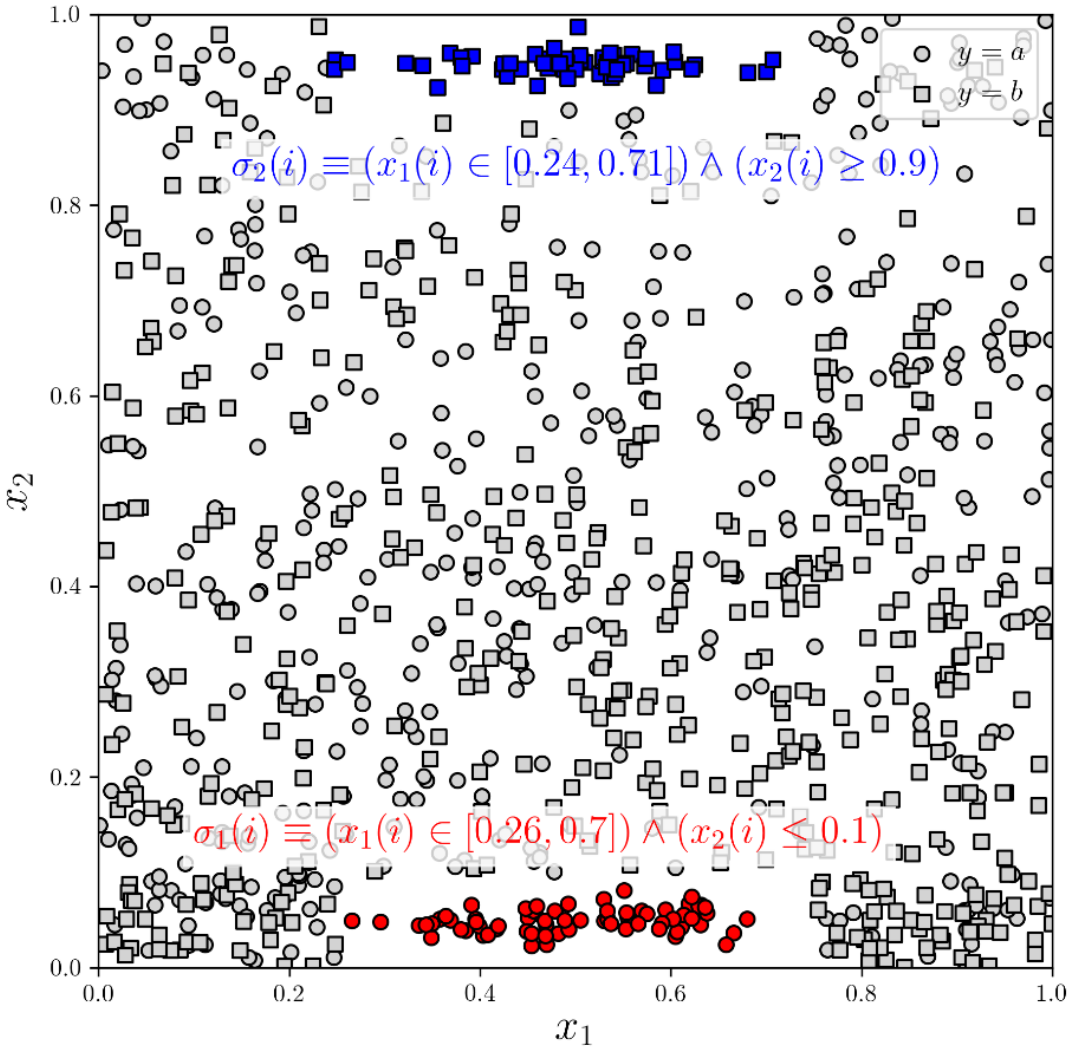
Decision trees



Subgroup discovery



Subgroup discovery: how it works



Descriptive features, $a_1, \ldots, a_m \in A$

e.g., d-band center, coordination number, atomic radii

Target features $y_1, \ldots, y_n \in Y$

e.g., adsorbate binding energy

Basic selectors, $c_1, \ldots, c_k \in C \rightarrow$ {false, true}

e.g., Is the d-band center low?; Is the atom coordination number < 3?

Find *selector* $\sigma = c_1(\cdot) \wedge \cdots \wedge c_l(\cdot)$

that maximizes *quality* $q = \left(\frac{|\text{ext}(\sigma)|}{|P|}\right)^{\alpha} u(Y_\sigma)^{1-\alpha}$

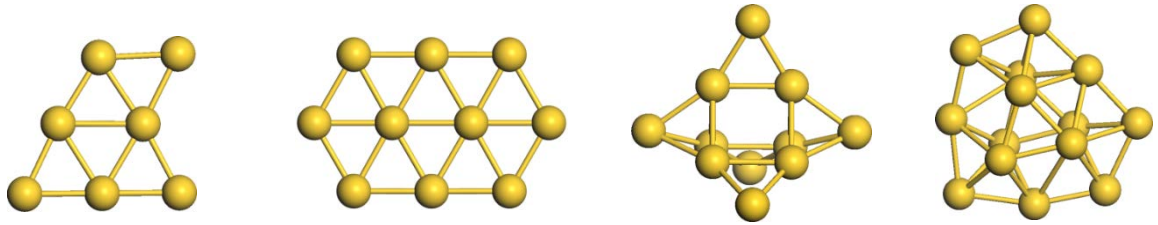
$\frac{|\text{ext}(\sigma)|}{|P|}$ is the coverage of points where σ is true

$u(Y_\sigma)$ is the utility function (optimization criteria)

Two tutorial applications of subgroup discovery presented

1. Gas Phase Gold Clusters (Au_5-Au_{14})

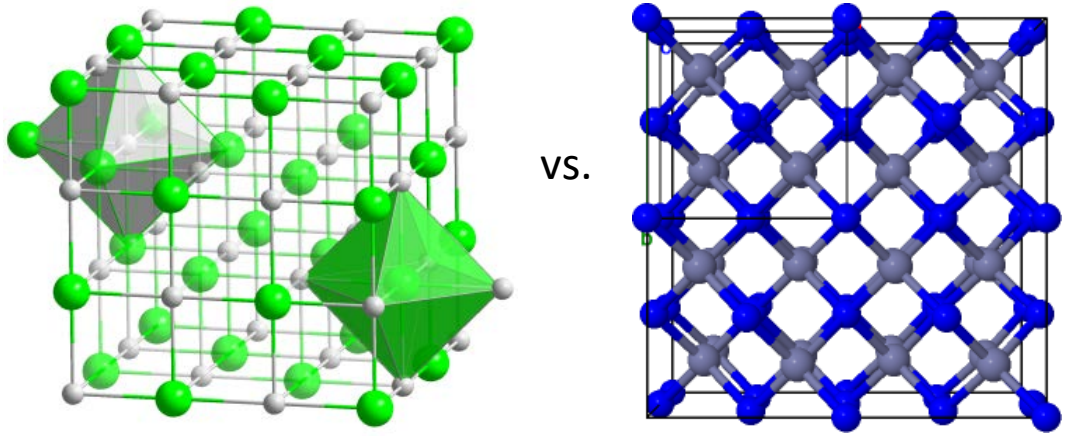
Display interesting catalytic properties



24,400 gold cluster configurations in total

2. Classification and description of 82 Octet Binary Semiconductors

Rocksalt vs. Zincblende

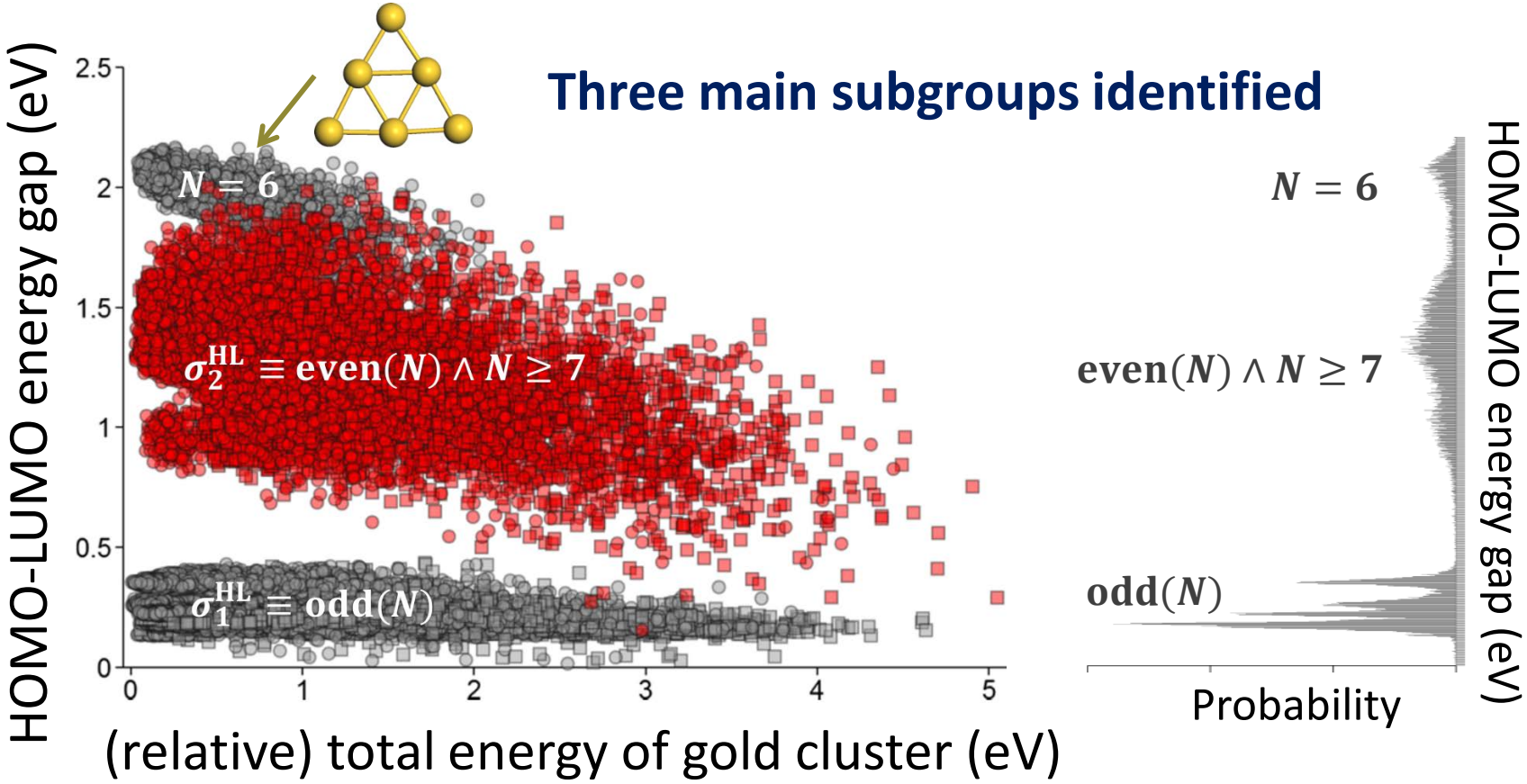


Rediscover simple insight about HOMO-LUMO gap

24,400 gold cluster configurations (Au_5-Au_{14}) in the gas phase

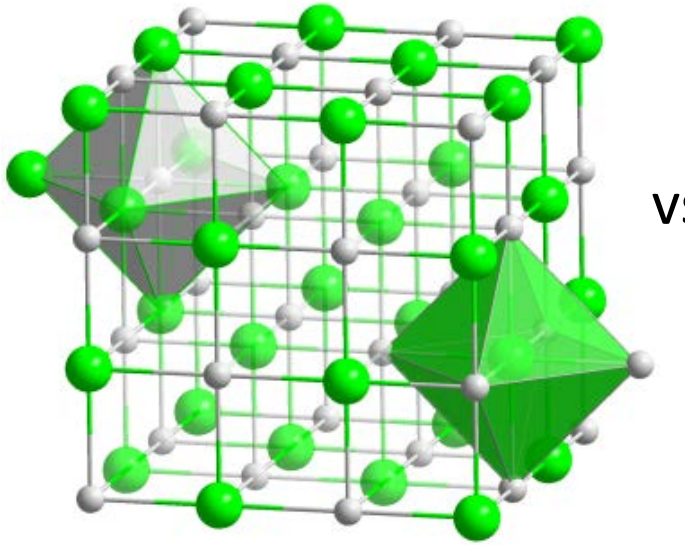
Choose **target property**
HOMO-LUMO energy gap

Choose **variation reduction utility function**
$u(Y') = (\mathrm{std}(Y) - \mathrm{std}(Y'))/\mathrm{std}(Y)$



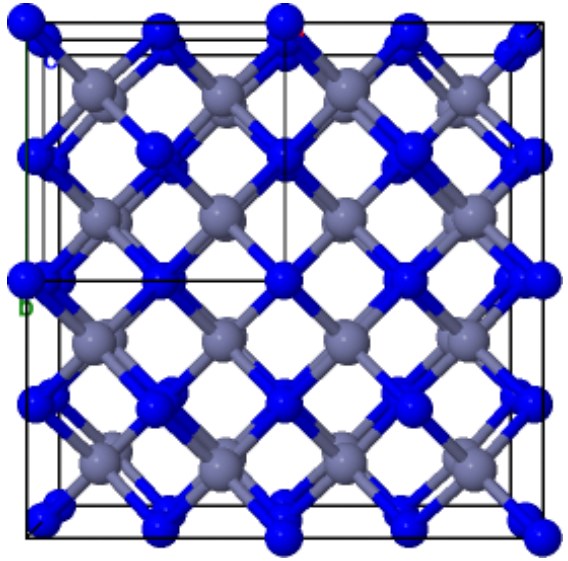
Find descriptors that predict crystal structures for the 82 octet AB-type materials

Rocksalt



vs.

Zincblende



Target property

$\text{sign}(E_{\text{rocksalt}} - E_{\text{zincblende}})$

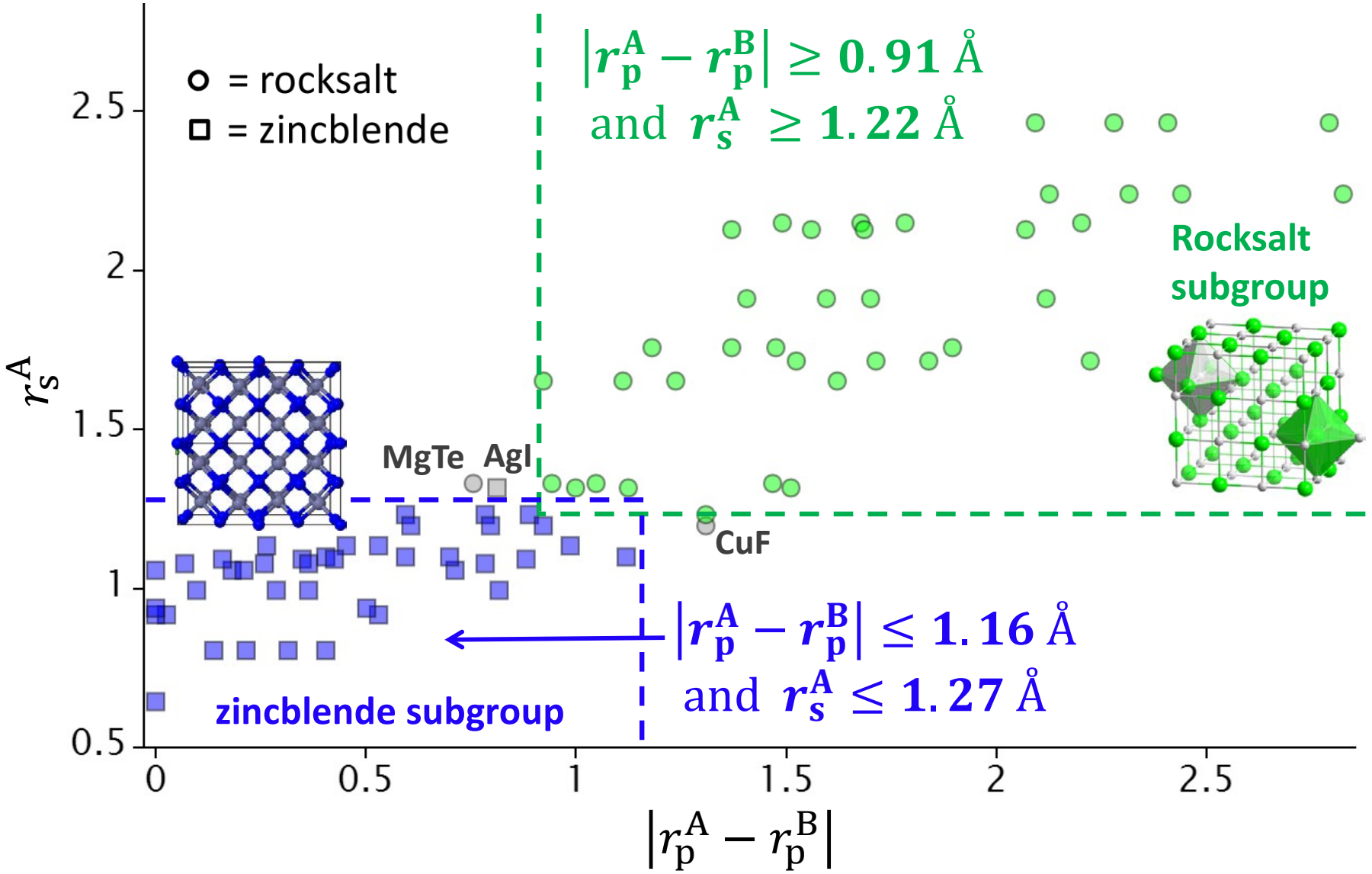
Input candidate descriptors into subgroup discovery from DFT calculations

- Radii of *s*, *p*, *d* orbitals of free atoms
- Electron affinity
- Ionization potential

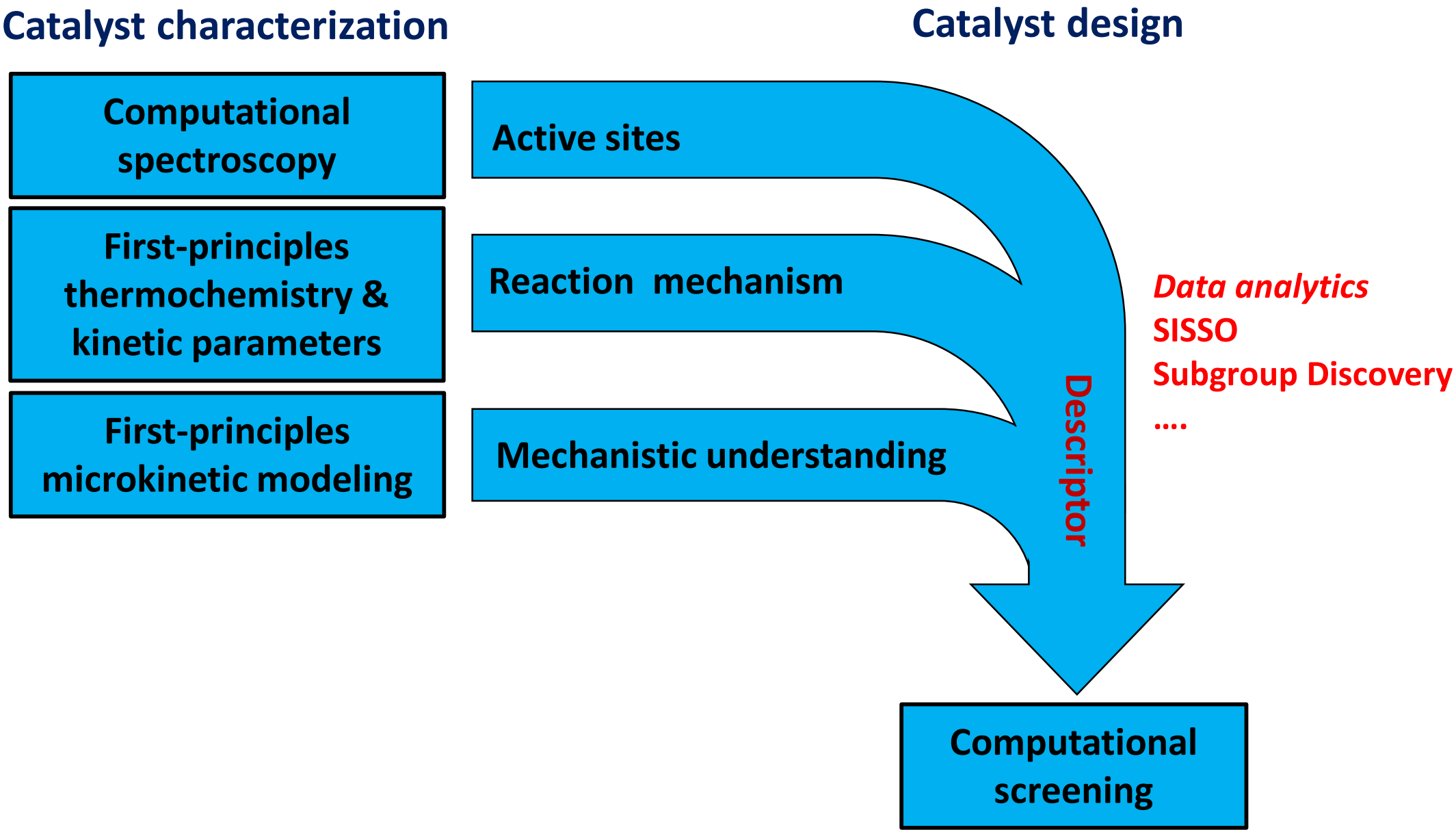
….and others

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

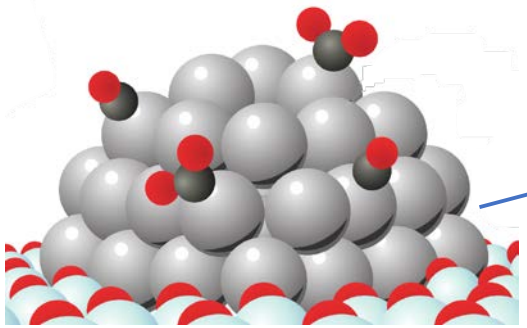
Subgroup discovery classifies 79 of the 82 compounds using a two-dimensional descriptor



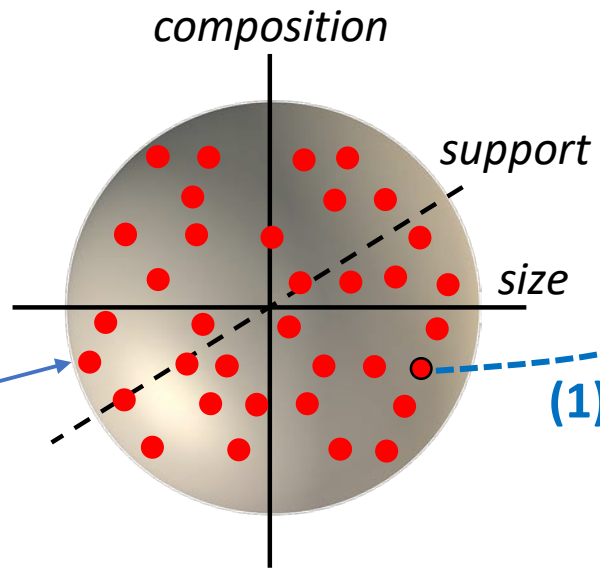
Looking ahead: Can we use data analytics to extract catalyst descriptors?



Machine learning for catalyst design and discovery

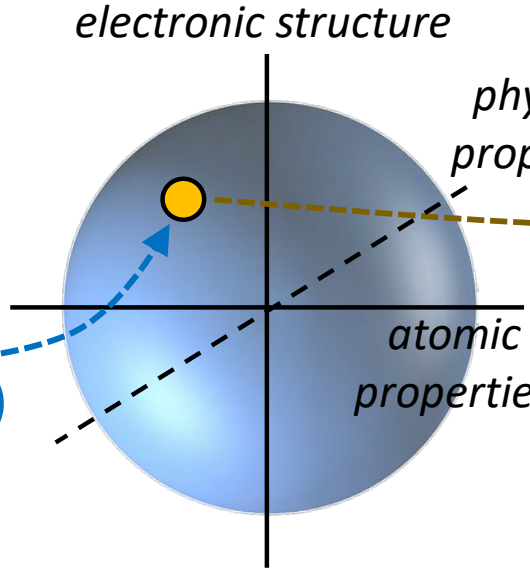


Heterogeneous Catalyst



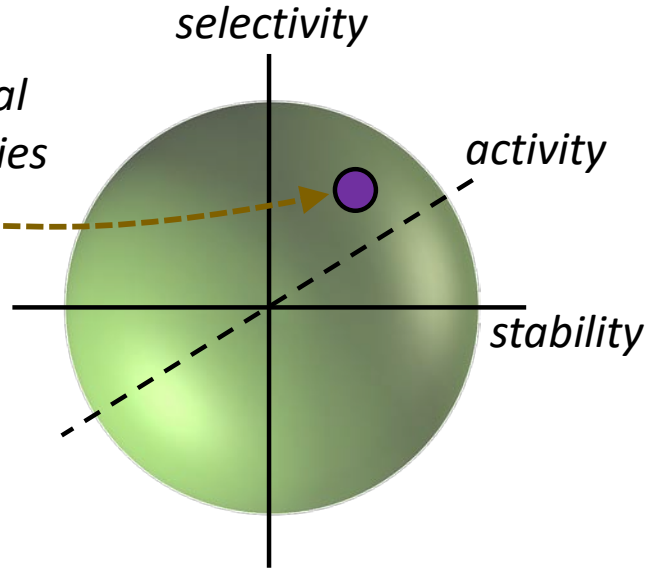
Catalyst Space

(1)



Feature Space

(2)



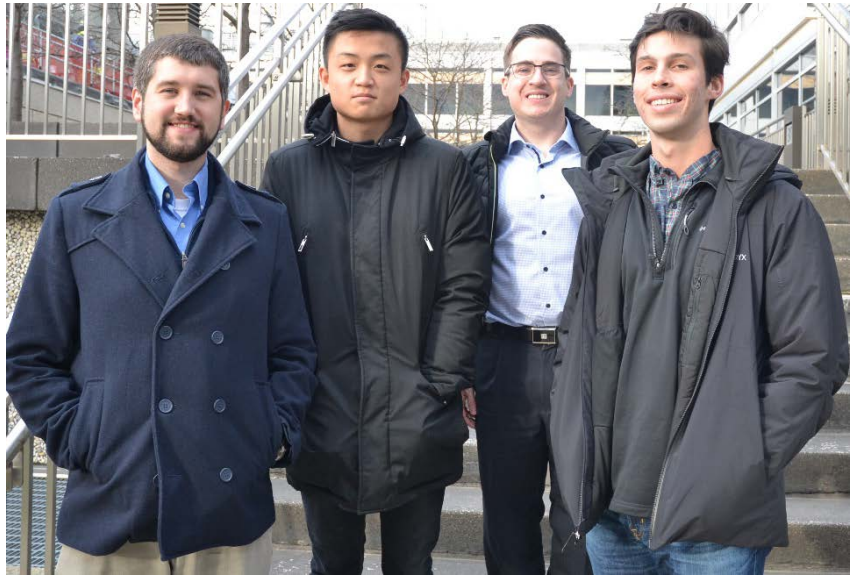
Figures of Merit

Machine learning
e.g., use subgroup discovery and compressed sensing
→ relate features to Figures of merit

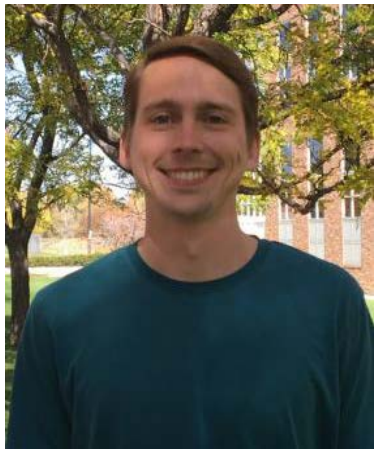
Acknowledgements

CHEMICAL ENGINEERING
UNIVERSITY OF MICHIGAN

Goldsmith Lab
Materials
Modeling
Data Science
Descriptors
Catalysis



Subgroup discovery
and SISSO tutorials are online
https://www.nomad-coe.eu/



Christopher Bartel

Mario Boley
Runhai Ouyang
Christopher Sutton
Jilles Vreeken
Matthias Scheffler
Luca M. Ghiringhelli
Charles Musgrave