Patterns, Correlations, and Discovery of Descriptors in Big Data of Materials

Bryan R. Goldsmith

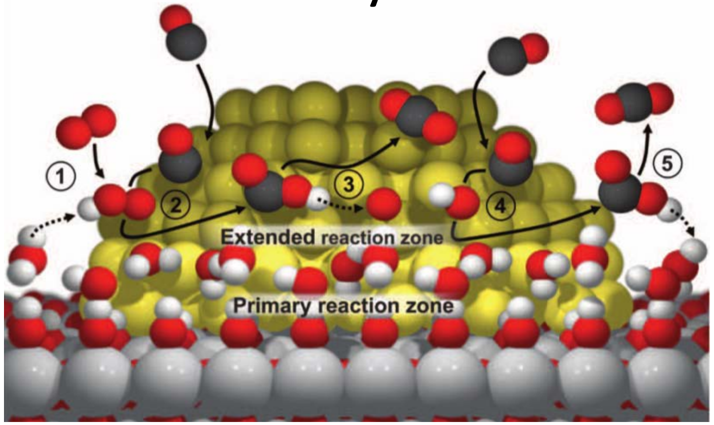
NOMAD

Fritz-Haber-Institut der Max-Planck-Gesellschaft
Theory Department

Alexander von Humboldt
Stiftung/Foundation

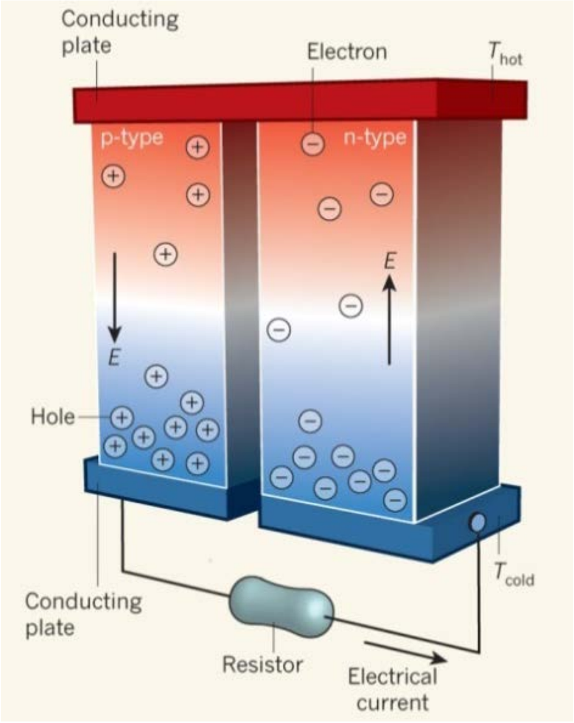
Designing materials requires an understanding of the mechanisms underlying a materials function

Catalysts

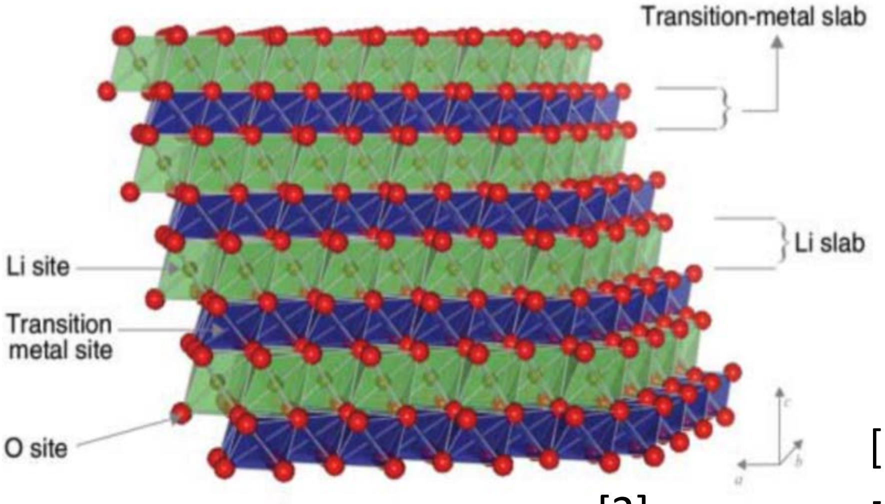


[1]

Thermoelectrics



Batteries



[2]

[1] J. Saavedra *et al. Science* 345, 1599 (2014)

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

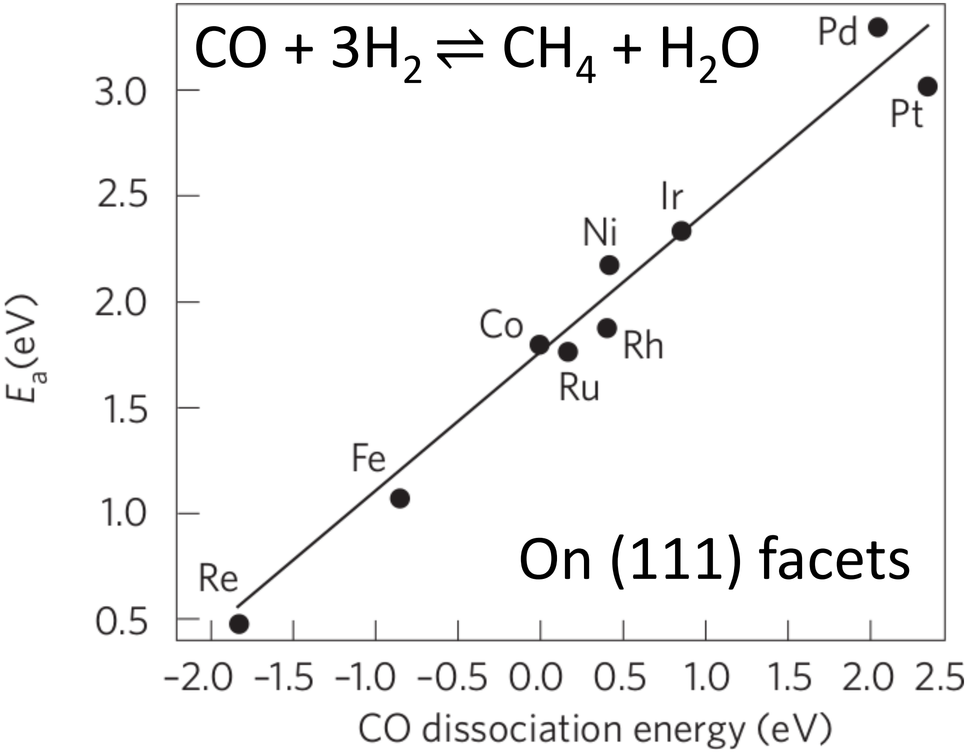
Identifying physically meaningful *descriptors* that describe materials properties is critical

Descriptor → Property

Descriptor = function(atomic or material features)

Predict new materials

Increase material understanding



J. K. Nørskov *et al. Nat. Chem.* 1, 37 (2009)

The development of data-analytic tools can facilitate the discovery of descriptors

This talk will focus on two different data-analytics tools to find descriptors of materials

1. Compressed sensing to find interpretable descriptors

Application 1: octet binary semiconductors

2. Subgroup discovery to find local patterns and their descriptions

Application 1: gold clusters in the gas phase (sizes 5-14 atoms)

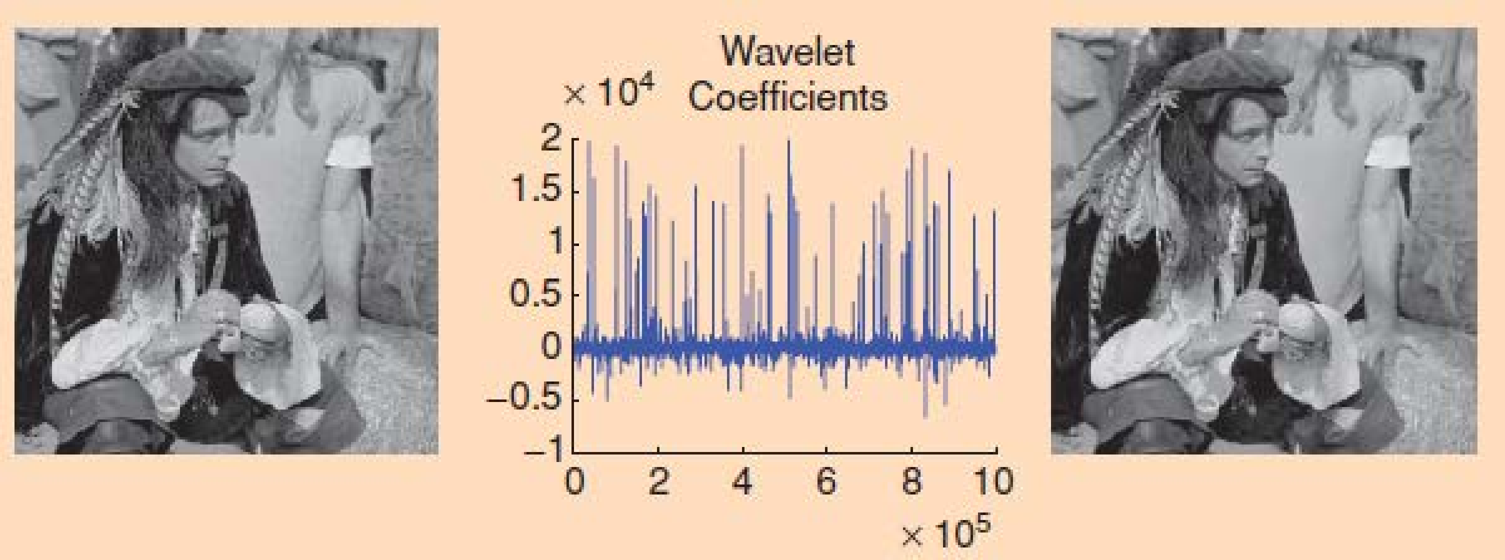
Application 2: octet binary semiconductors

Compressed sensing allows the construction of sparse models with high accuracy

Original image

Sparse
in the basis set

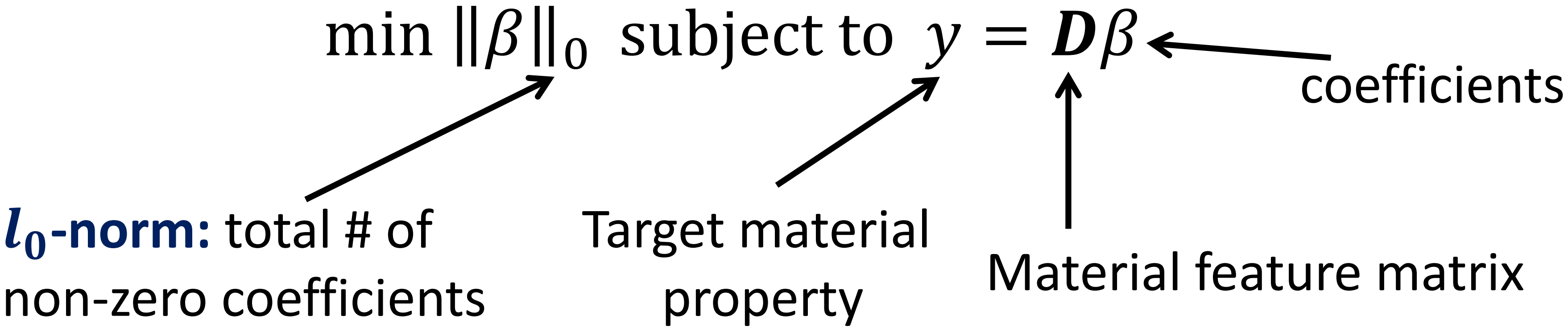
Recovered with 10% measurements



Emmanuel Candès, Terence Tao, and David Donoho

Compressed sensing allows the construction of sparse models with high accuracy

l_0-norm minimization



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

coefficients

l_0-norm: total # of non-zero coefficients

Target material property

Material feature matrix

Emmanuel Candès, Terence Tao, and David Donoho

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

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

l_1-norm:
Sum of absolute value of coefficients

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

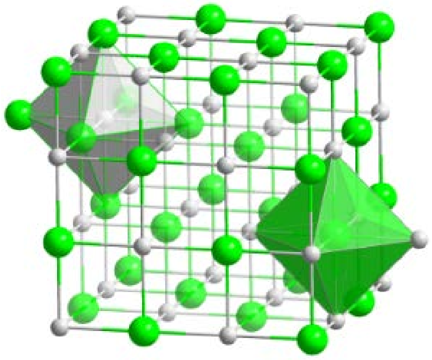
Root mean squared error

Regularization parameter

Find descriptors that predict the crystal structure energy differences between 82 octet binary compounds

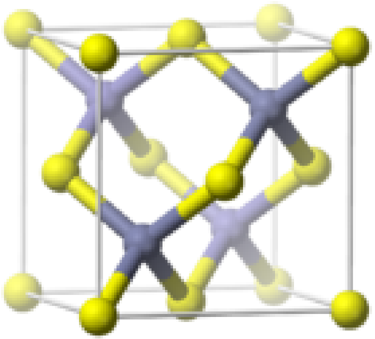
Target property = $E_{\text{rocksalt}} - E_{\text{zincblende}}$

Rocksalt (RS)



vs.

Zincblende (ZB)



Predict energy differences between 0.01 – 1.5 eV

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

1. Build a large feature space of candidate descriptors

Start from atomic properties (primary features)

$$\mathrm{IP(A)}\ \mathrm{EA(A)}\ \mathrm{IP(B)}\ \mathrm{EA(B)}$$
$$\mathrm{H(A)}\ \mathrm{L(A)}\ \mathrm{H(B)}\ \mathrm{L(B)}$$

$$r_s(\mathrm{A})\ r_p(\mathrm{A})\ r_d(\mathrm{A})$$
$$r_s(\mathrm{B})\ r_p(\mathrm{B})\ r_d(\mathrm{B})$$

Build up feature matrix ***D*** of 4500 compound features

operator set: {+,-,exp,^{2},÷,×}

$$r_s(\mathrm{A})^2, (r_p(\mathrm{A}) + r_s(\mathrm{A}))^2$$
$$\exp(r_s(\mathrm{A})), \exp(r_p(\mathrm{A}) \pm r_s(\mathrm{A}))$$

$$|\mathrm{IP(A)} \pm \mathrm{IP(B)}|$$
$$|\mathrm{L(B)} \pm \mathrm{H(A)}|$$
$$|r_p(\mathrm{A}) \pm r_s(\mathrm{A})|$$

2. Feature selection using two step scheme: **LASSO+l_0**

Step 1: **LASSO**

Reduce the feature space to find candidate set of possible descriptors (e.g., 4500 → 50 features)

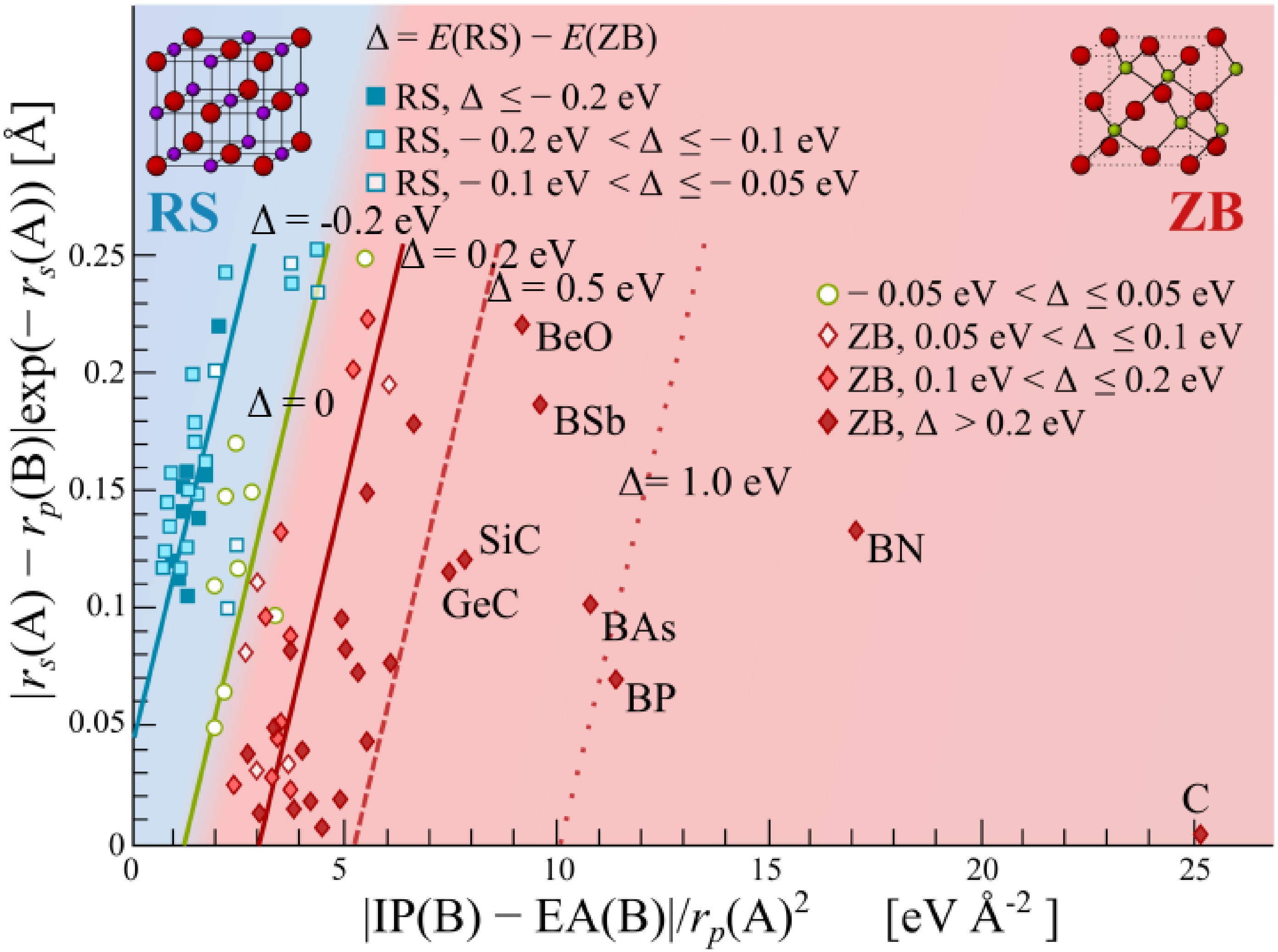
$$\hat{\beta}_{LASSO}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{2} \|y - \boldsymbol{D}\beta\|_2^2 + \lambda \|\beta\|_1 \right)$$

Step 2: **l_0 minimization**

Find the 'best' *n*-dimensional model from the candidate set $\boldsymbol{D}`$

$$\hat{\beta}_{l_0}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{2} \|y - \boldsymbol{D}`\beta\|_2^2 + \lambda \|\beta\|_0 \right)$$

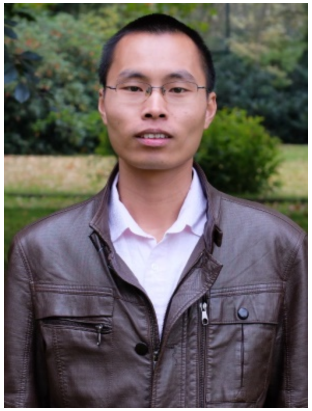
A highly predictive two-dimensional descriptor is found for the 82 octet binary semiconductors



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

LASSO has stability issues for huge feature space of correlated features

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



Runhai Ouyang, Luca M. Ghiringhelli,
Emhre Ahmetcik, Matthias Scheffler

R. Ouyang, E. Ahmetcik, L. M. Ghiringhelli, M. Scheffler, *unpublished*

1. Systematically construct a huge feature space (**one trillion**)

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

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

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

Sure independent screening

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

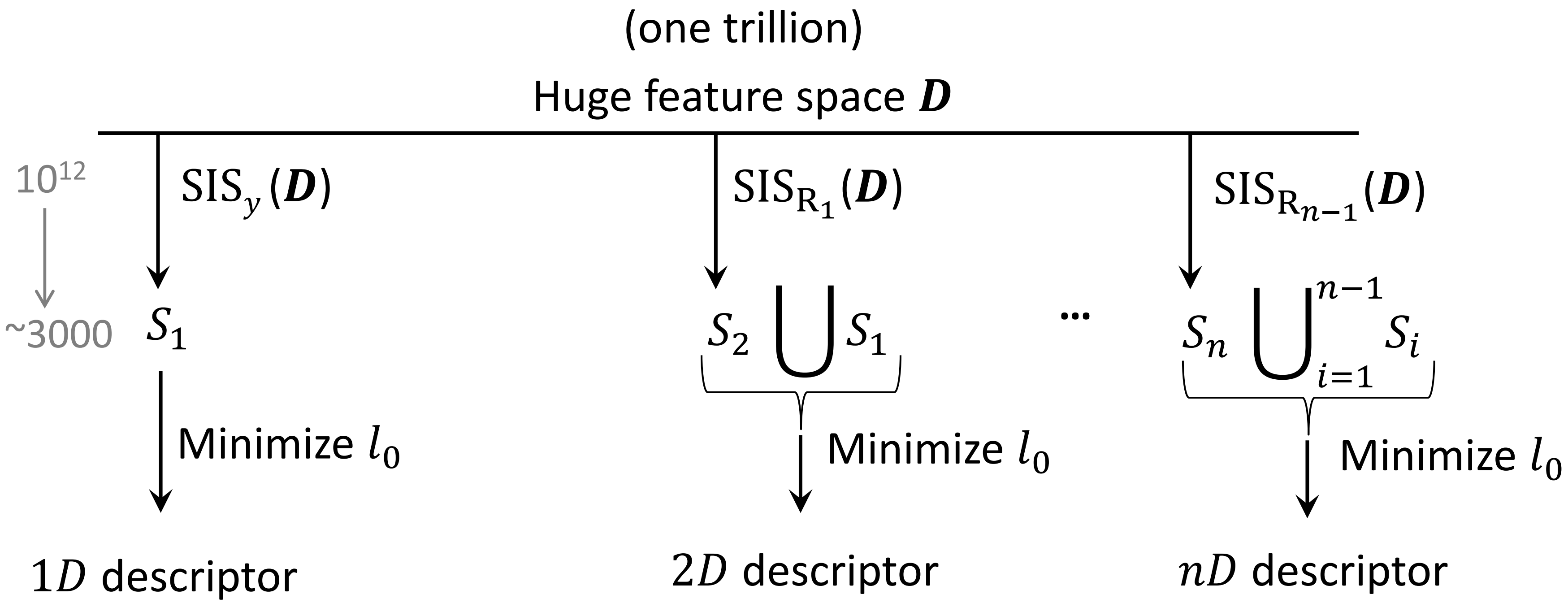
Results in a subspace of *N* features that are most correlated with the property of interest

$\boldsymbol{D}$: matrix of the feature space

y: target property ($E_{\text{rocksalt}} – E_{\text{zincblende}}$)

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

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



(one trillion)

Huge feature space $\boldsymbol{D}$

10^{12}

~3000

$\mathrm{SIS}_y(\boldsymbol{D})$

S_1

Minimize l_0

$1D$ descriptor

$\mathrm{SIS}_{\mathrm{R}_1}(\boldsymbol{D})$

$\underbrace{S_2 \bigcup S_1}$

Minimize l_0

$2D$ descriptor

…

$\mathrm{SIS}_{\mathrm{R}_{n-1}}(\boldsymbol{D})$

$\underbrace{S_n \bigcup_{i=1}^{n-1} S_i}$

Minimize l_0

nD descriptor

SIS = sure independent screening

S_i = feature subspace

y = target property

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

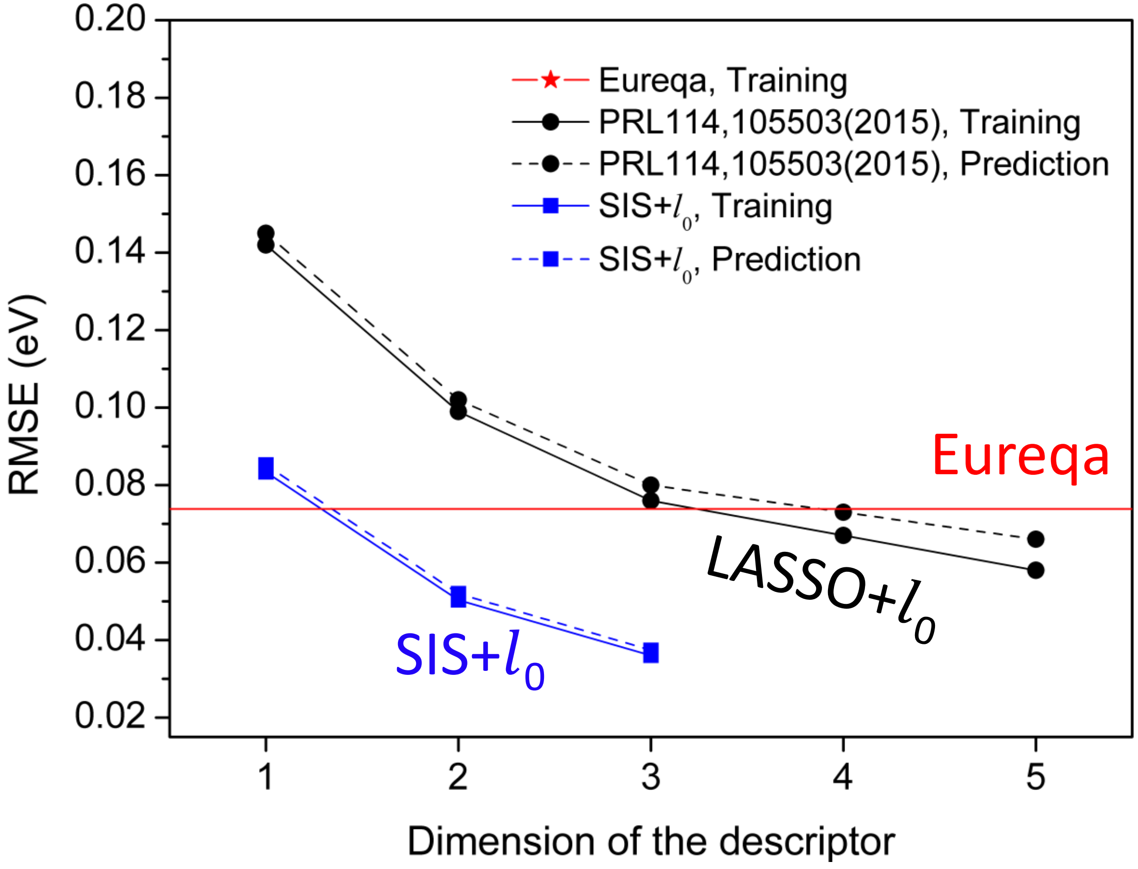
R. Ouyang, E. Ahmetcik, L. M. Ghiringhelli, M. Scheffler, *unpublished*

Accurate and interpretable descriptors are found from a **one trillion** dimension feature space

Target property = $E_{\mathrm{rocksalt}} - E_{\mathrm{zincblende}}$

1D descriptor:

$$\frac{|r_s(B)-r_d(A)| * r_s(B)}{r_d(A) * (r_p(A)^3 + r_p(B)^3)}$$



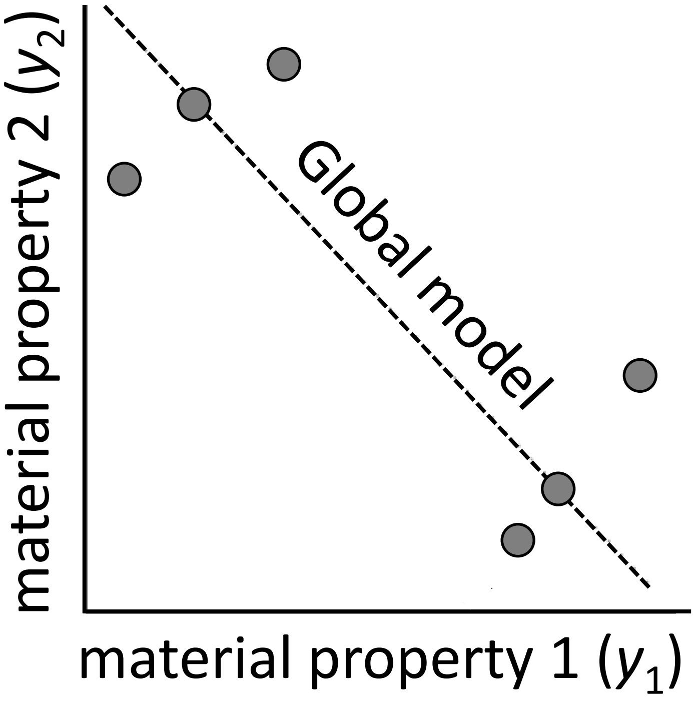
R. Ouyang, E. Ahmetcik, L. M. Ghiringhelli, M. Scheffler, *unpublished*

What if data is heterogeneous and the global model is too complex to be interpretable?

Underlying mechanisms can change across materials

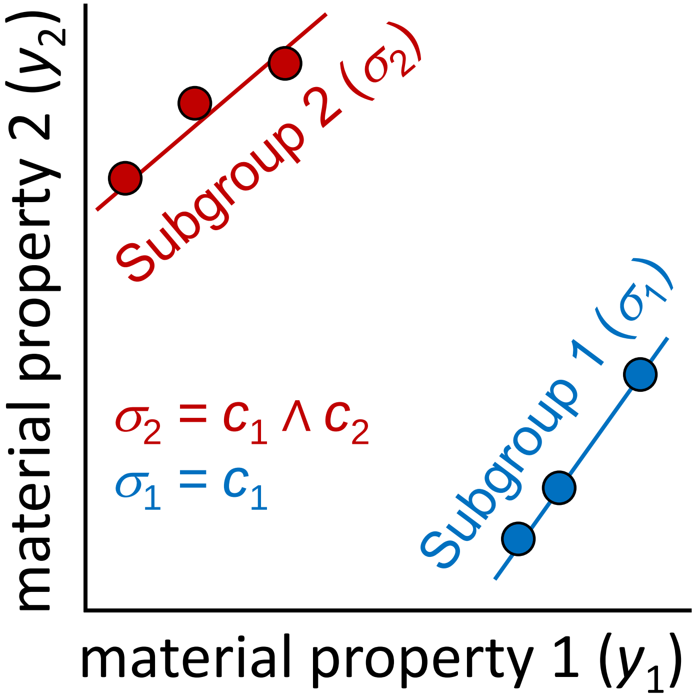
Relations between subsets of data may be important

Goal: find *physically interpretable local models* of a target property in materials data


material property 2 (y_2)
material property 1 (y_1)
Global model

Subgroup discovery: find physically interpretable local models of a target property in materials data

Mario Boley, me,
Luca Ghiringhelli

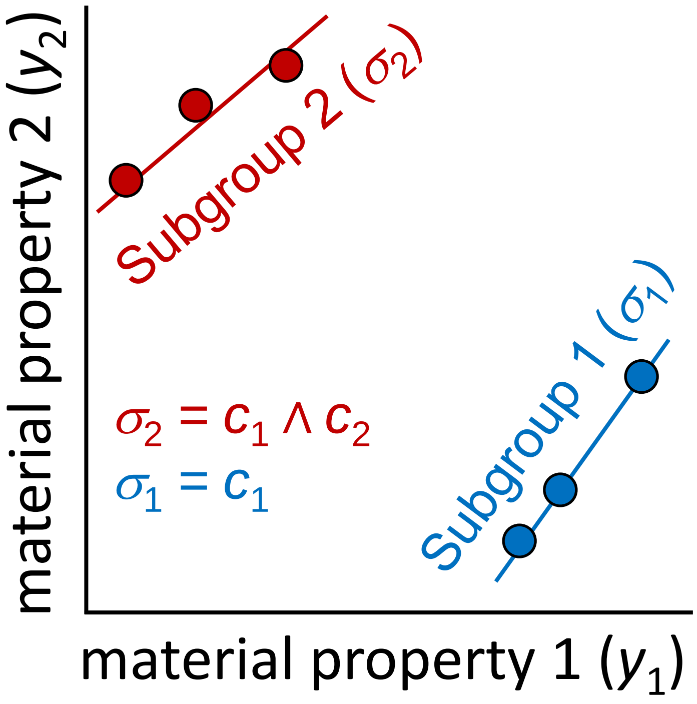


Subgroup 2 (σ_2)
Subgroup 1 (σ_1)
$\sigma_2 = c_1 \wedge c_2$
$\sigma_1 = c_1$
material property 2 (y_2)
material property 1 (y_1)

M. Atzmueller, *WIREs Data Min. Knowl. Discov. 5* (2015)

W. Duivesteijn, A. J. Feelders, A. Knobbe, *Data Min. Knowl. Discov. 30*, 47 (2016)

Subgroup discovery: how it works


Subgroup 2 (σ_2)
Subgroup 1 (σ_1)
$\sigma_2 = c_1 \wedge c_2$
$\sigma_1 = c_1$
material property 2 (y_2)
material property 1 (y_1)

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

e.g., energy, bonding topology, number of atoms

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

e.g., HOMO-LUMO energy gap

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

e.g., Is there an even number of atoms?

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

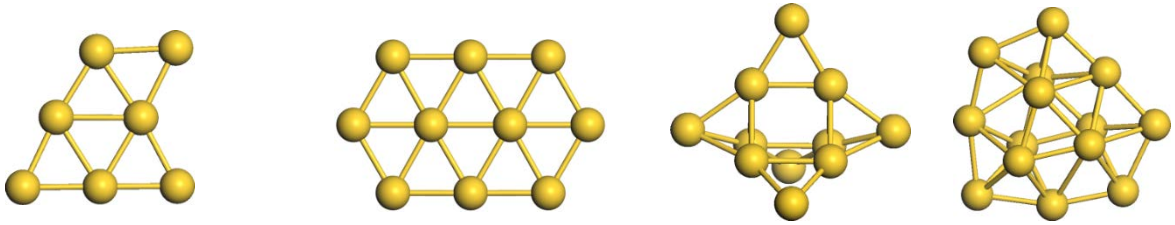
$u(Y_\sigma)$ is the utility function (optimization criteria)

Subgroup discovery (SGD) implemented
by Mario Boley using Creedo and realKD library

Two applications of subgroup discovery

1. Gas Phase Gold Clusters (size 5-14)

Display interesting optical, chemical, and electronic properties



ca. 25,000 gold cluster configurations in total

2. Classification of 82 Octet Binary Semiconductors

Rocksalt   vs.   Zincblende

B. R. Goldsmith, M. Boley, J. Vreeken, L. M. Ghiringhelli, M. Scheffler, *unpublished*

Rediscover simple insight about HOMO-LUMO gap

25,000 gold cluster configurations (sizes 5-14) in the gas phase

Choose **target property**

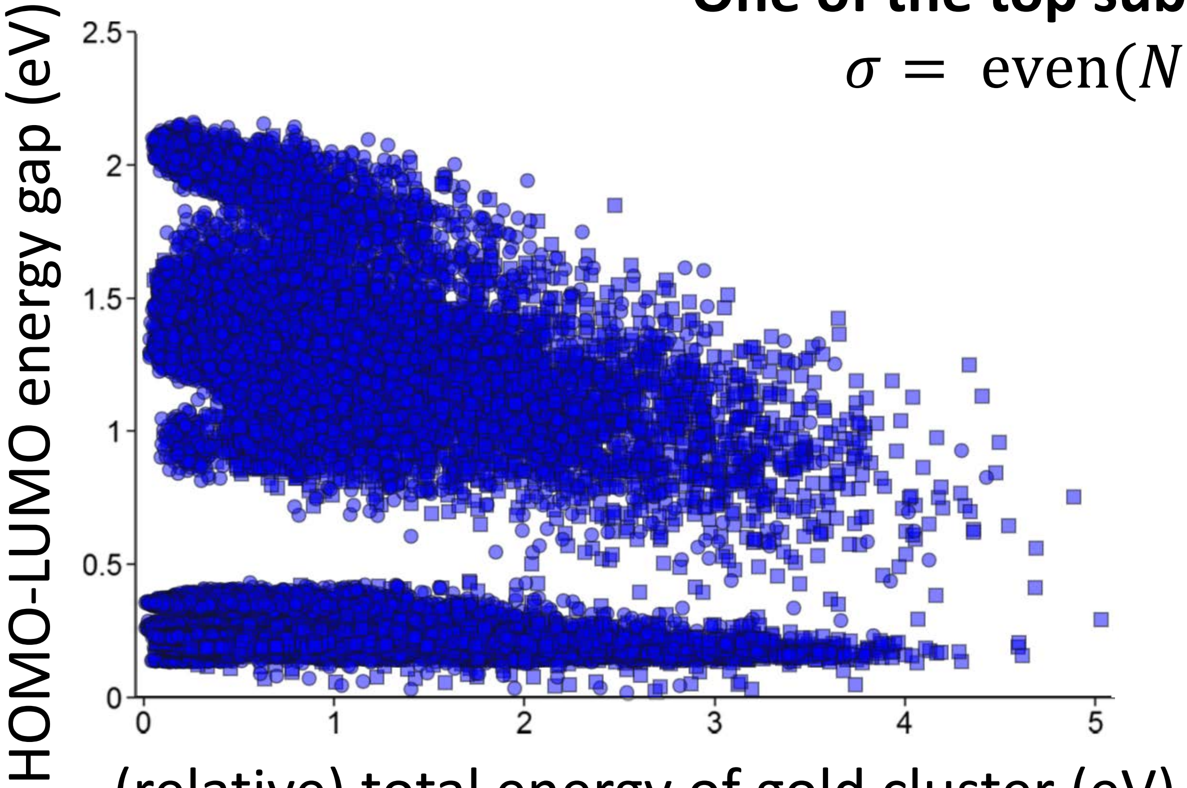
HOMO-LUMO energy gap

Choose **variation reduction utility function**

$$u(Y') = (\mathrm{std}(Y) - \mathrm{std}(Y'))/\mathrm{std}(Y)$$

One of the top subgroup selectors found

$$\sigma = \text{ even}(N) \wedge N \geq 7$$



Rediscover simple insight about HOMO-LUMO gap

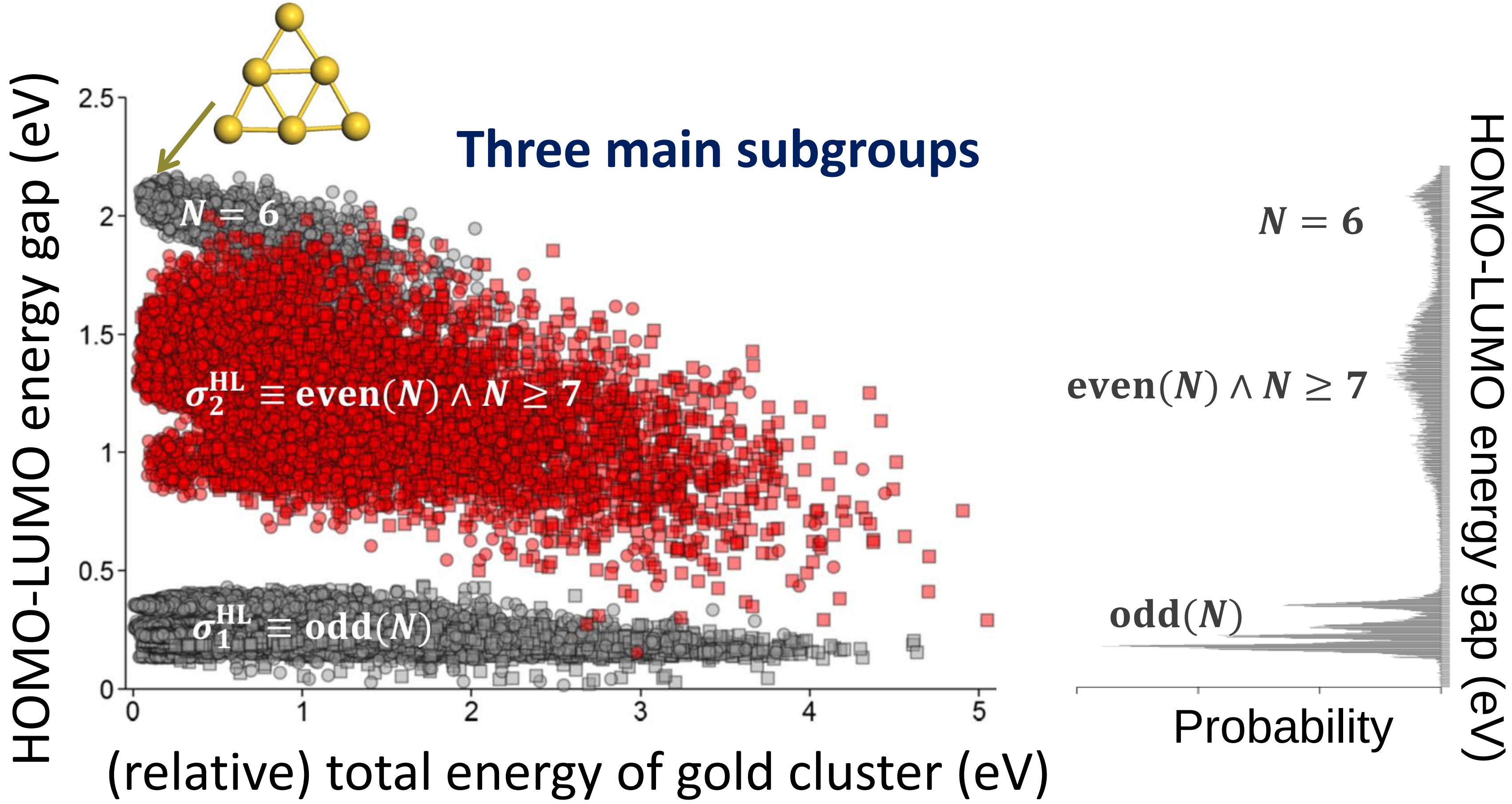
25,000 gold cluster configurations (sizes 5-14) in the gas phase

Choose **target property**

HOMO-LUMO energy gap

Choose **variation reduction utility function**

$$u(Y') = (\mathrm{std}(Y) - \mathrm{std}(Y'))/\mathrm{std}(Y)$$



Three main subgroups

Equilibrium state has maximum electronic hardness at 0 K

Electronic hardness = resistance to electron density deformation

$$\left.\frac{\partial\mu}{\partial N}\right|_{v} \approx \frac{1}{2}(E_{\mathrm{LUMO}} - E_{\mathrm{HOMO}})$$

Can electronic hardness be a descriptor for cluster isomer stability?

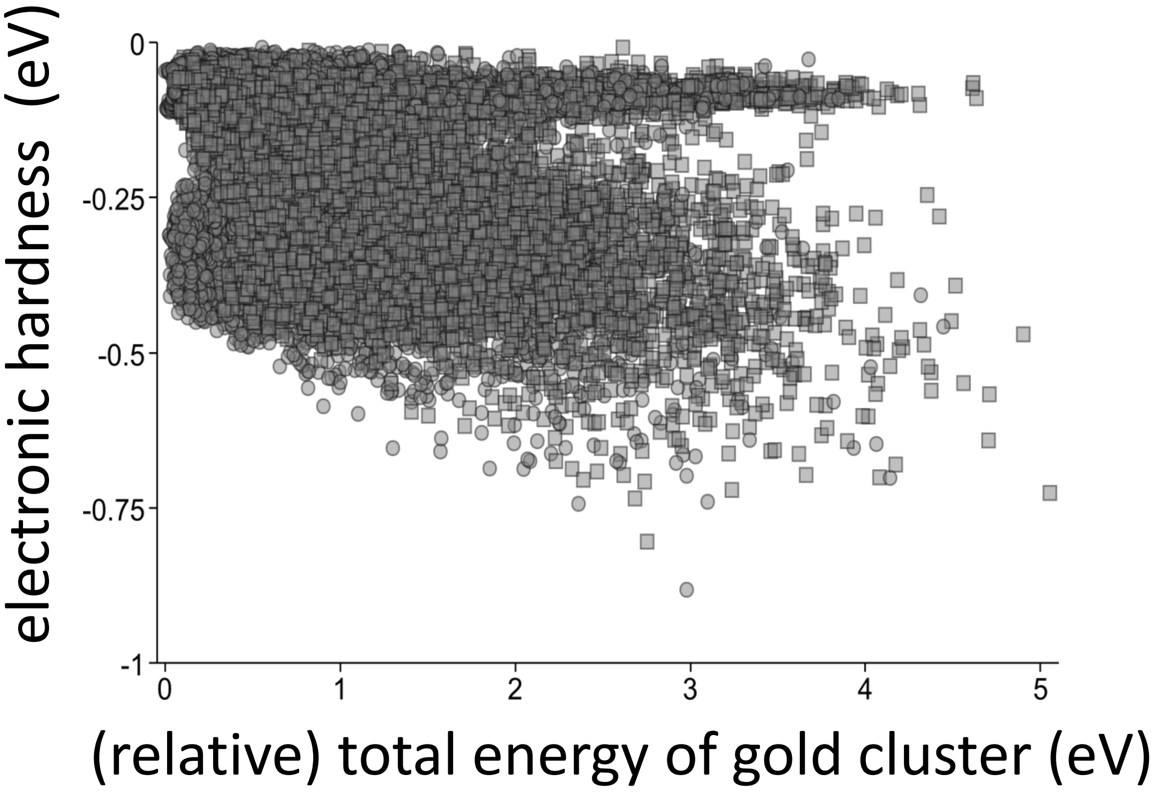
Choose target property

electronic hardness and formation energy

Find a linear model: $u(Y_1', Y'_2)$

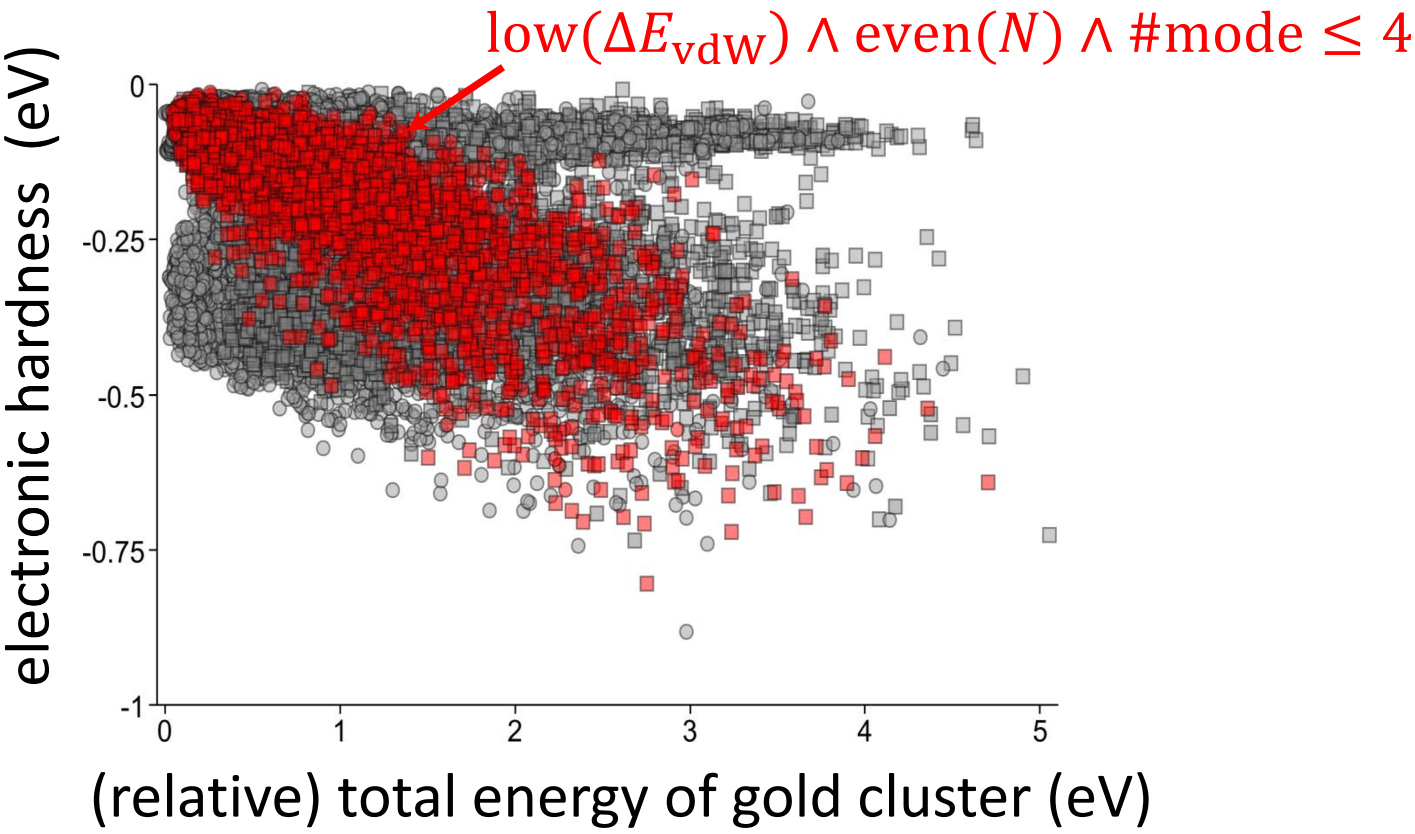
Top subgroup selector found

$$\mathrm{low}(\Delta E_{\mathrm{vdW}}) \wedge \mathrm{even}(N) \wedge \#\mathrm{mode} \leq 4$$


electronic hardness (eV)
0
-0.25
-0.5
-0.75
-1
0
1
2
3
4
5
(relative) total energy of gold cluster (eV)

Electronic hardness can be a qualitative descriptor for stability beyond the ground state (among the subgroup)

Subgroup population = 30% of even sized clusters

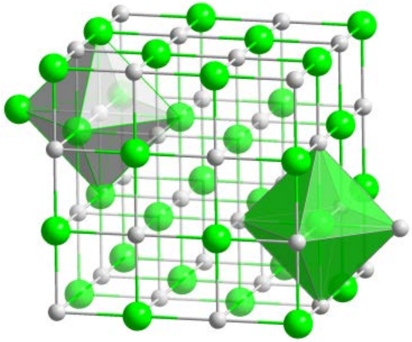


Can subgroup discovery find local models for the 82 octet binary materials that describe ZB and RS?

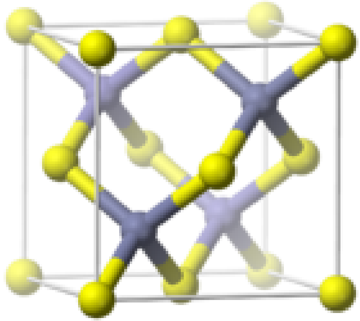
$$\mathit{Target} = \mathrm{sign}\,(E_{\mathrm{rocksalt}} - E_{\mathrm{zincblende}})$$

Subgroup 1 ↓ Rocksalt (RS)

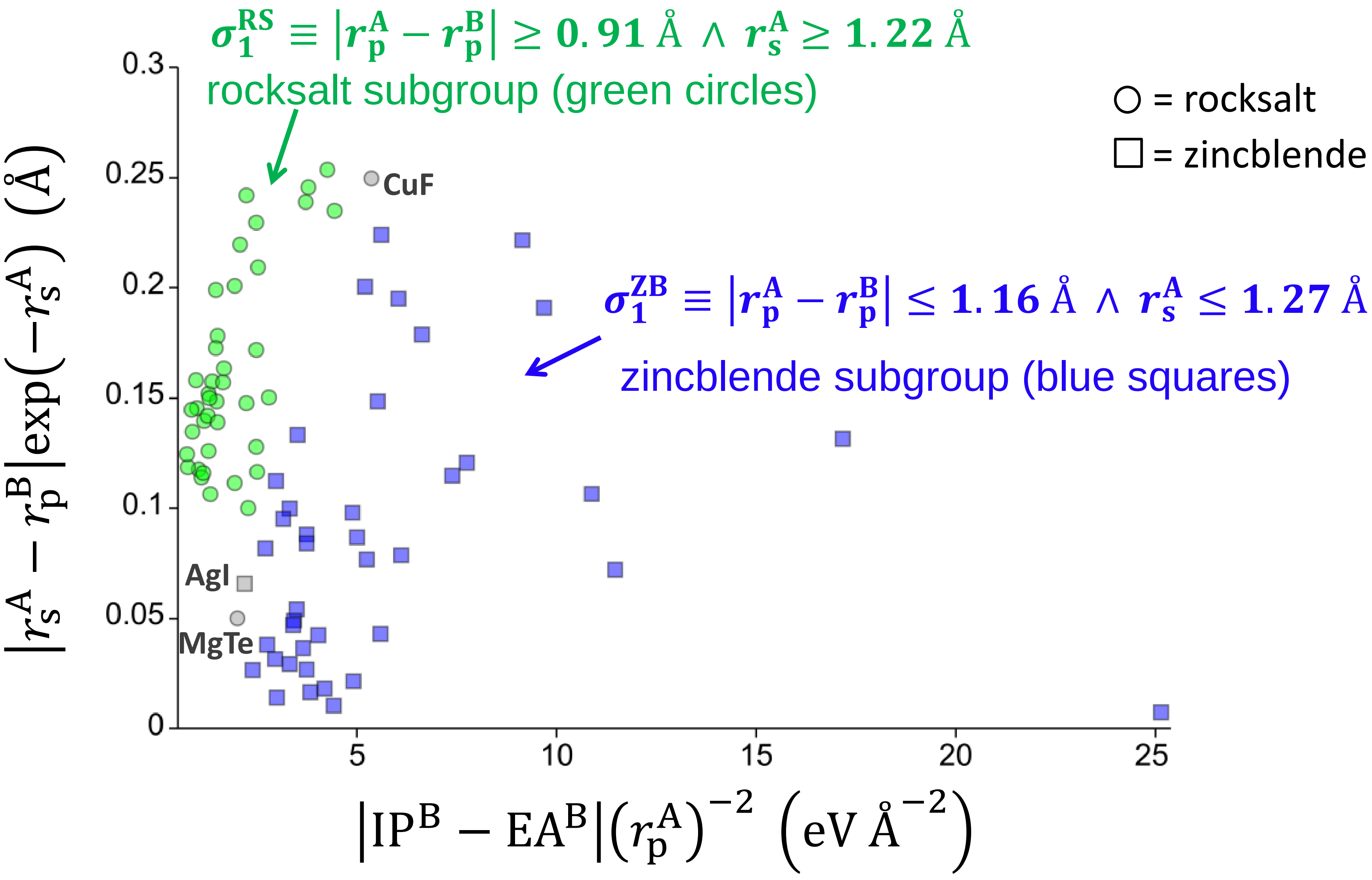
Subgroup 2 ↓ Zincblende (ZB)



vs.



Subgroup discovery classifies 79 of 82 octet binary semiconductors using a 2D descriptor



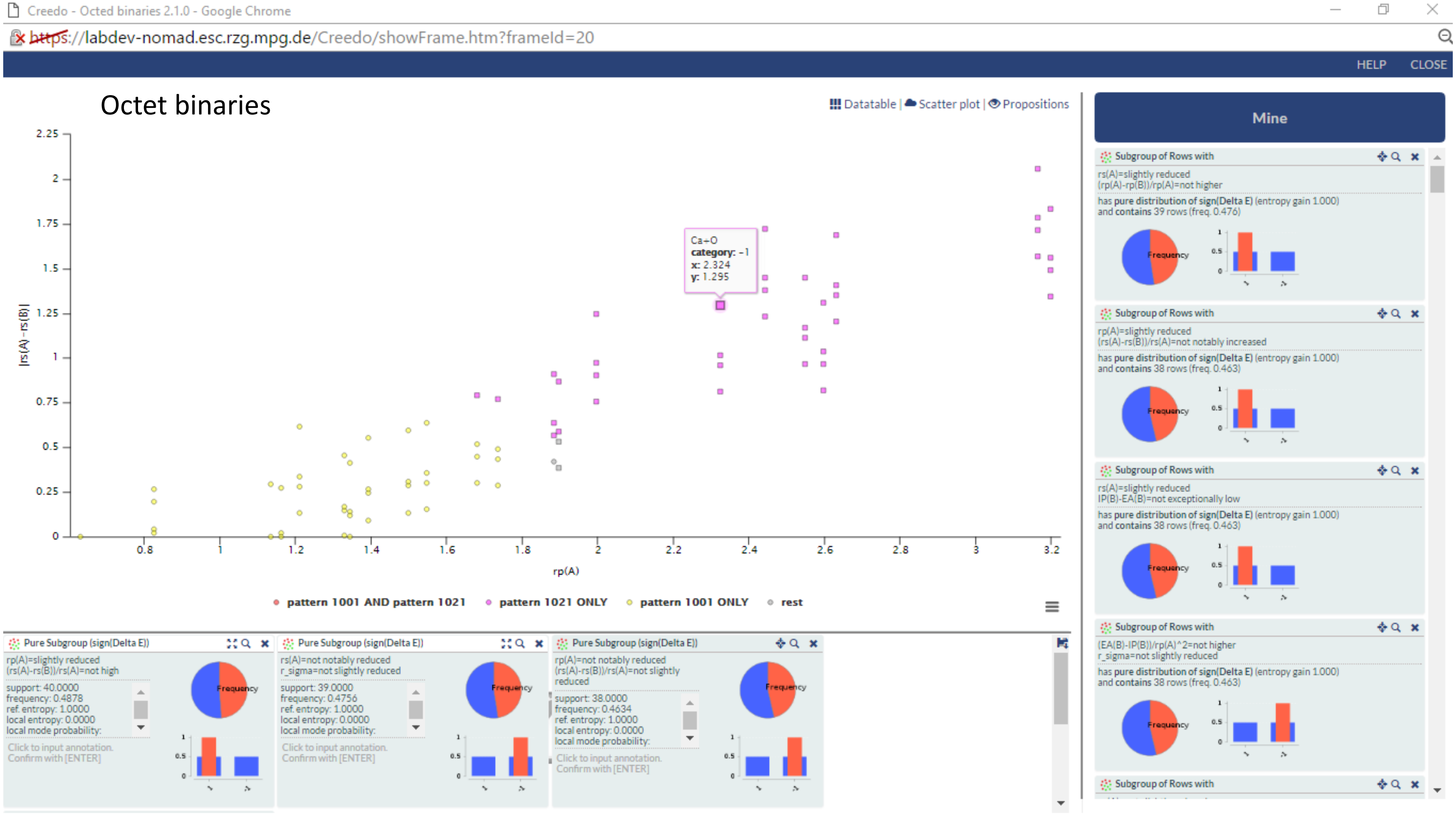
$\boldsymbol{\sigma_1^{RS} \equiv \left|r_p^A - r_p^B\right| \geq 0.91}$ Å $\wedge$ $\boldsymbol{r_s^A \geq 1.22}$ Å

rocksalt subgroup (green circles)

$\boldsymbol{\sigma_1^{ZB} \equiv \left|r_p^A - r_p^B\right| \leq 1.16}$ Å $\wedge$ $\boldsymbol{r_s^A \leq 1.27}$ Å

zincblende subgroup (blue squares)

We have made subgroup discovery, among other data analytic tools, available online as interactive tutorials



Pure Subgroup (sign(Delta E))

rp(A)=slightly reduced
(rs(A)-rs(B))/rs(A)=not high

support: 40.0000
frequency: 0.4878
ref. entropy: 1.0000
local entropy: 0.0000
local mode probability:

Click to input annotation.
Confirm with [ENTER]

Pure Subgroup (sign(Delta E))

rs(A)=not notably reduced
r_sigma=not slightly reduced

support: 39.0000
frequency: 0.4756
ref. entropy: 1.0000
local entropy: 0.0000
local mode probability:

Click to input annotation.
Confirm with [ENTER]

Pure Subgroup (sign(Delta E))

rp(A)=not notably reduced
(rs(A)-rs(B))/rs(A)=not slightly reduced

support: 38.0000
frequency: 0.4634
ref. entropy: 1.0000
local entropy: 0.0000
local mode probability:

Click to input annotation.
Confirm with [ENTER]

Subgroup of Rows with

rs(A)=slightly reduced
(rp(A)-rp(B))/rp(A)=not higher

has **pure distribution of sign(Delta E)** (entropy gain 1.000)
and **contains** 39 rows (freq. 0.476)

Subgroup of Rows with

rp(A)=slightly reduced
(rs(A)-rs(B))/rs(A)=not notably increased

has **pure distribution of sign(Delta E)** (entropy gain 1.000)
and **contains** 38 rows (freq. 0.463)

Subgroup of Rows with

rs(A)=slightly reduced
IP(B)-EA(B)=not exceptionally low

has **pure distribution of sign(Delta E)** (entropy gain 1.000)
and **contains** 38 rows (freq. 0.463)

Subgroup of Rows with

(EA(B)-IP(B))/rp(A)^2=not higher
r_sigma=not slightly reduced

has **pure distribution of sign(Delta E)** (entropy gain 1.000)
and **contains** 38 rows (freq. 0.463)

Subgroup of Rows with

Subgroup discovery (SGD) was implemented by Mario Boley using Creedo and realKD library, www.realkd.org

Big-data analytics tools for materials science are being developed by the NOMAD team

Compressed sensing (LASSO, l_0, LASSO+l_0)

Sure independent screening+l_0

Subgroup discovery

The question of causal inference is still open (advice is welcome!)

Bigger data and harder problems

Crystal structure maps

Transparent conducting oxides

…. and much more!

Subgroup discovery: B. Goldsmith, M. Boley, J. Vreeken, L. Ghiringhelli, M. Scheffler, *unpublished*

SIS+L0: R. Ouyang, E. Ahmetcik, L. Ghiringhelli, M. Scheffler, *unpublished*

LASSO+L0: L. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, *PRL 114*, 105503 (2015)

Acknowledgements

The entire NOMAD consortium

Special thanks:

Matthias Scheffler
Luca Ghiringhelli
Mario Boley
Jilles Vreeken
Runhai Ouyang
Christopher Sutton
Matthias Rupp
Claudia Draxl

NOMAD

Fritz Haber Institute of the Max Planck Society
Theory Department

Alexander von Humboldt
Stiftung/Foundation

mpi max planck institut informatik