NOMAD
NOVEL MATERIALS DISCOVERY

FAIRmat

MAX-PLANCK-GESELLSCHAFT

Identifying interpretable descriptors for materials properties with subgroup discovery and information theory

Luca M. Ghiringhelli
FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

Big Data Summer
A summer school of the BiGmax Network
Platja d'Aro, Spain, September 9 - 13, 2019

NOMAD
NOVEL MATERIALS DISCOVERY

FAIRmat

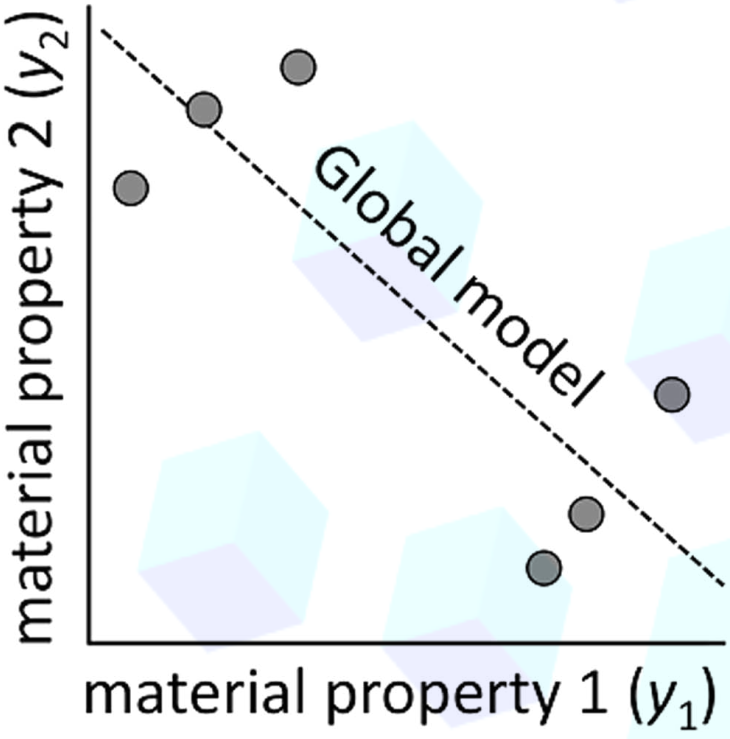
Subgroup discovery as research assistant: meta-learning and questioning old ideas

MAX-PLANCK-GESELLSCHAFT

Luca M. Ghiringhelli
FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

Big Data Summer
A summer school of the BiGmax Network
Platja d'Aro, Spain, September 9 - 13, 2019

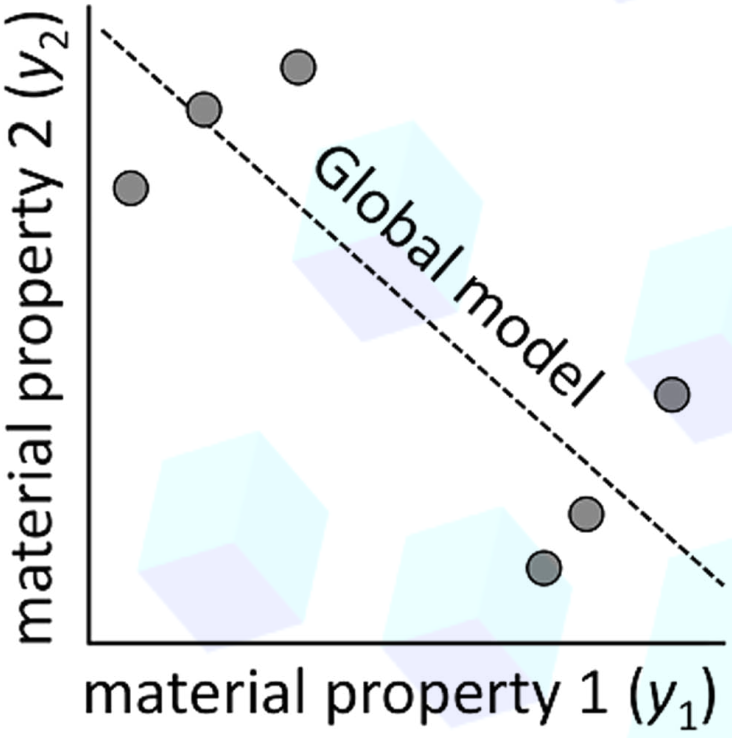
Global vs local learning: subgroup discovery

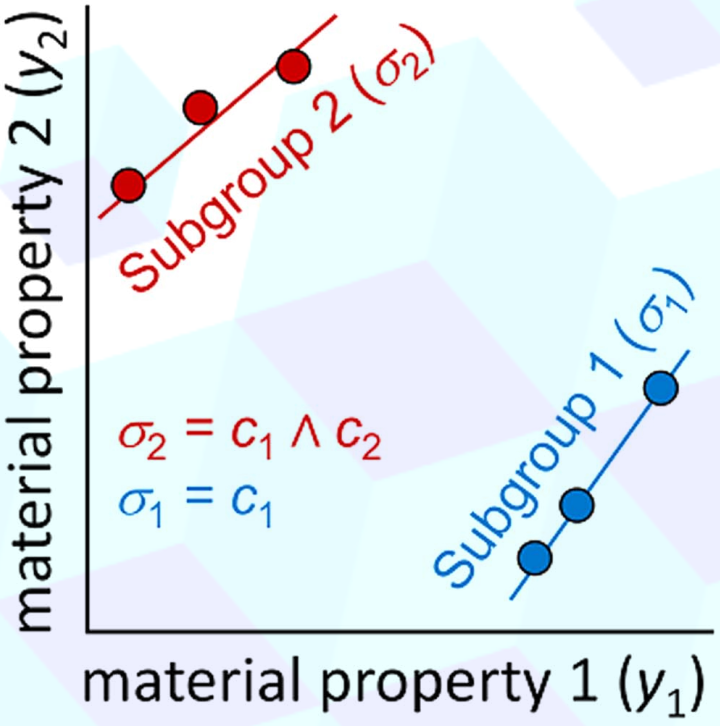


Boley, Goldsmith, LMG & Vreeken, Data Min. Knowl. Disc. **31**, 1391 (2017)
Goldsmith, Boley, Vreeken, Scheffler & LMG, New J. Phys. **19**, 013031 (2017)

Global vs local learning: subgroup discovery

NOVEL MATERIALS DISCOVERY





Boley, Goldsmith, LMG & Vreeken, Data Min. Knowl. Disc. **31**, 1391 (2017)
Goldsmith, Boley, Vreeken, Scheffler & LMG, New J. Phys. **19**, 013031 (2017)

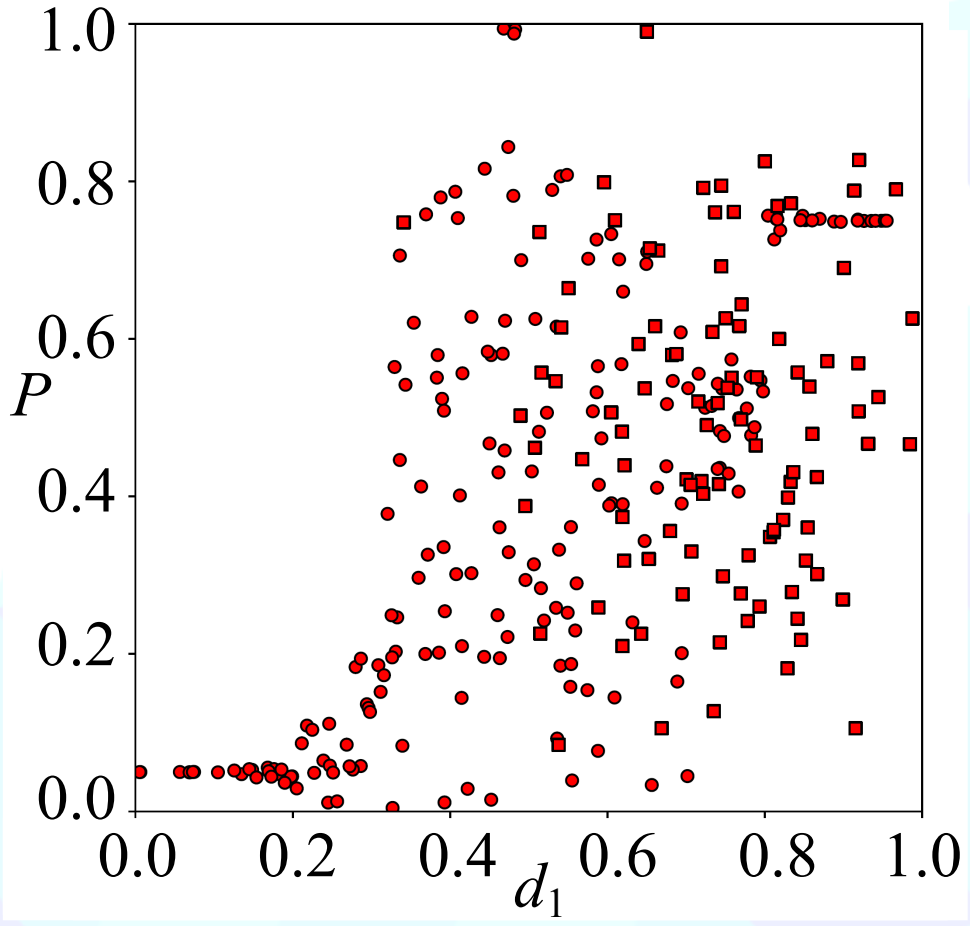
Global vs local learning: subgroup discovery

Ingredients:

Sample $S \subseteq$ population

Target property P_j

Features (descriptors) $\boldsymbol{d}_j$



Global vs local learning: subgroup discovery

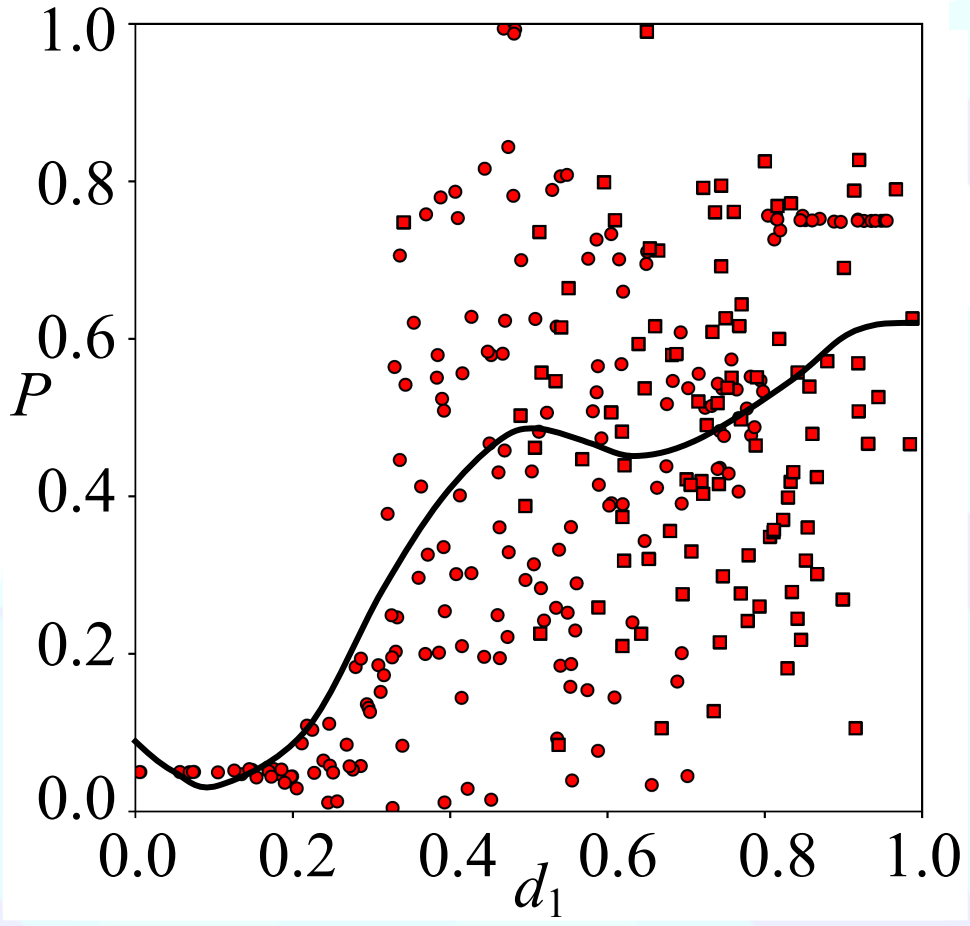
NOVEL MATERIALS DISCOVERY

Ingredients:

Sample $S \subseteq$ population

Target property P_j

Features (descriptors) $\boldsymbol{d}_j$



Global vs local learning: subgroup discovery

NOVEL MATERIALS DISCOVERY

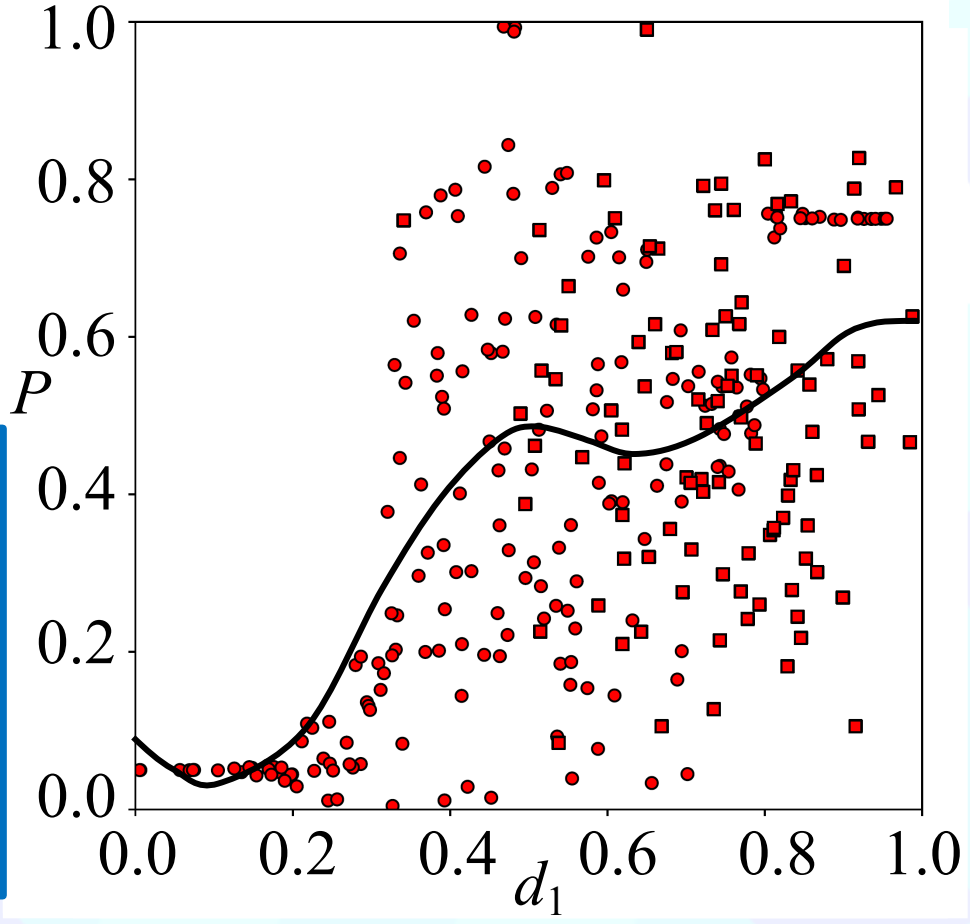
Ingredients:

Sample $S \subseteq$ population

Target property P_j

Features (descriptors) $\boldsymbol{d}_j$

A global model (fitted to the entire dataset) may hide or incorrectly describe the actuating physical mechanisms.



Global vs local learning: subgroup discovery

NOVEL MATERIALS DISCOVERY

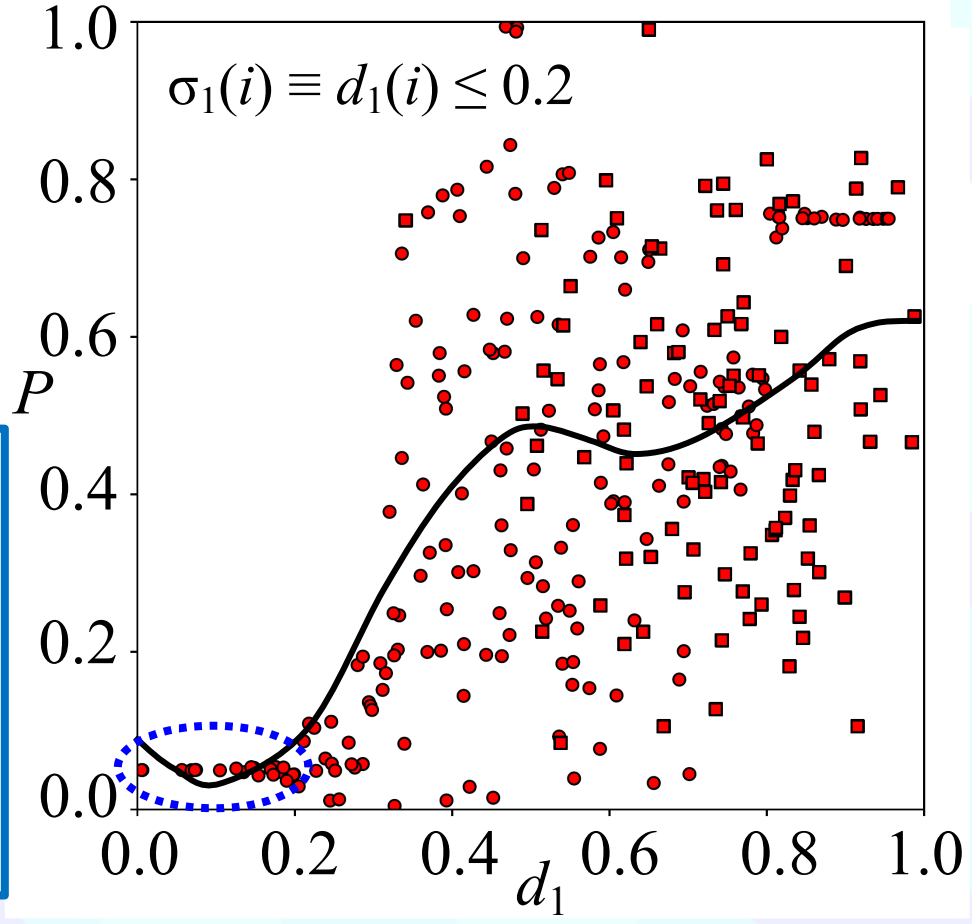
Ingredients:

Sample $S \subseteq$ population

Target property P_j

Features (descriptors) $\boldsymbol{d}_j$

A global model (fitted to the entire dataset) may hide or incorrectly describe the actuating physical mechanisms.



Global vs local learning: subgroup discovery

NOVEL MATERIALS DISCOVERY

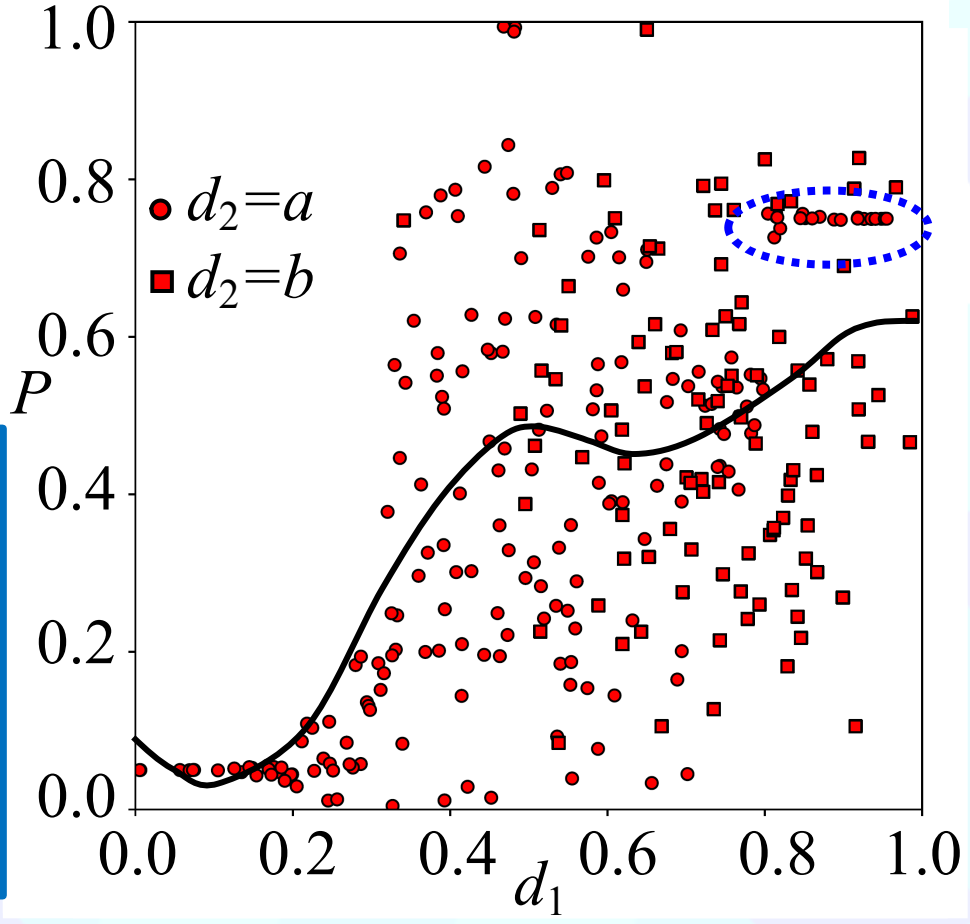
Ingredients:

Sample $S \subseteq$ population

Target property P_j

Features (descriptors) $\boldsymbol{d}_j$

A global model (fitted to the entire dataset) may hide or incorrectly describe the actuating physical mechanisms.



Global vs local learning: subgroup discovery

NOVEL MATERIALS DISCOVERY

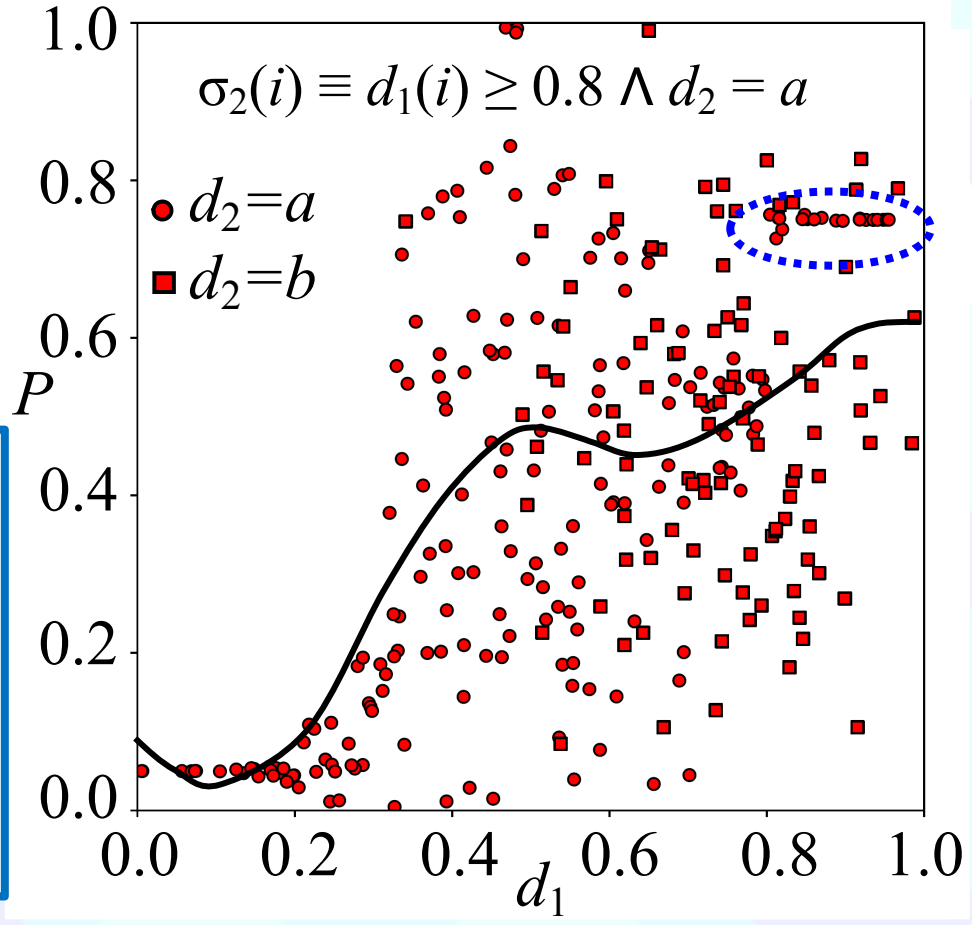
Ingredients:

Sample $S \subseteq$ population

Target property P_j

Features (descriptors) $\boldsymbol{d}_j$

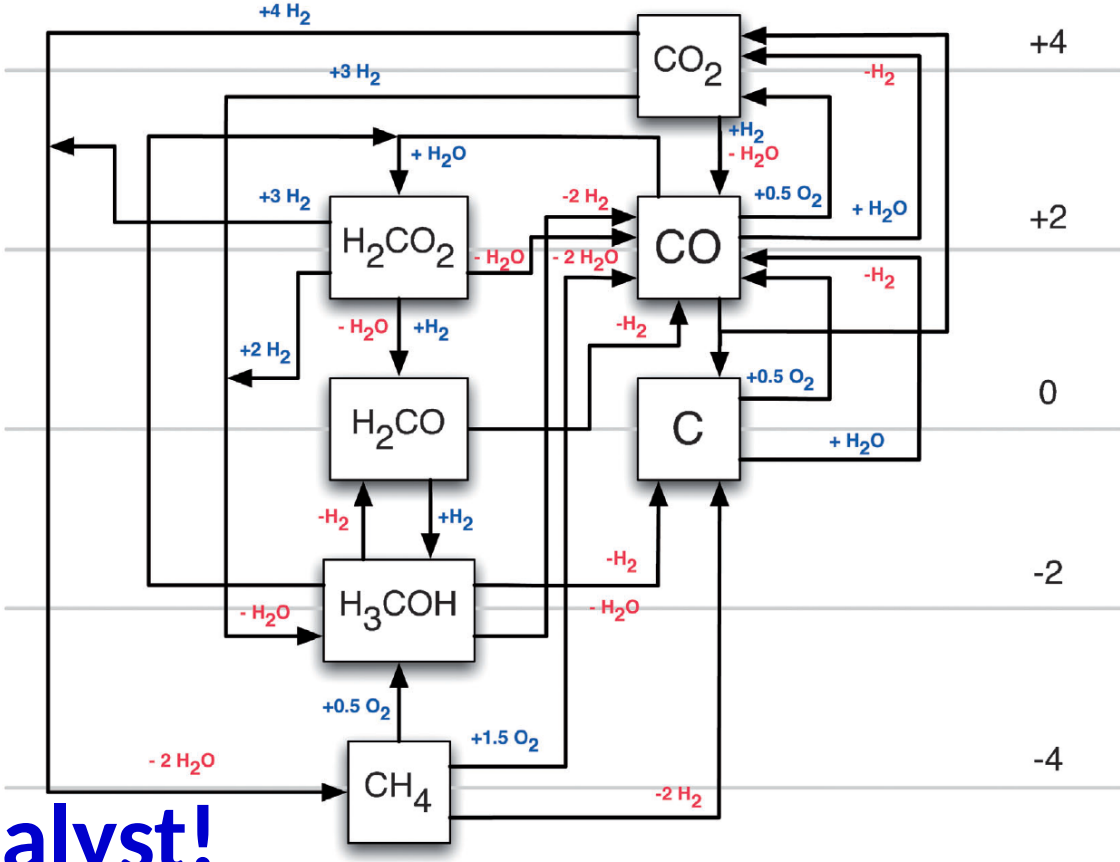
A global model (fitted to the entire dataset) may hide or incorrectly describe the actuating physical mechanisms.



NOMAD

NOVEL MATERIALS DISCOVERY

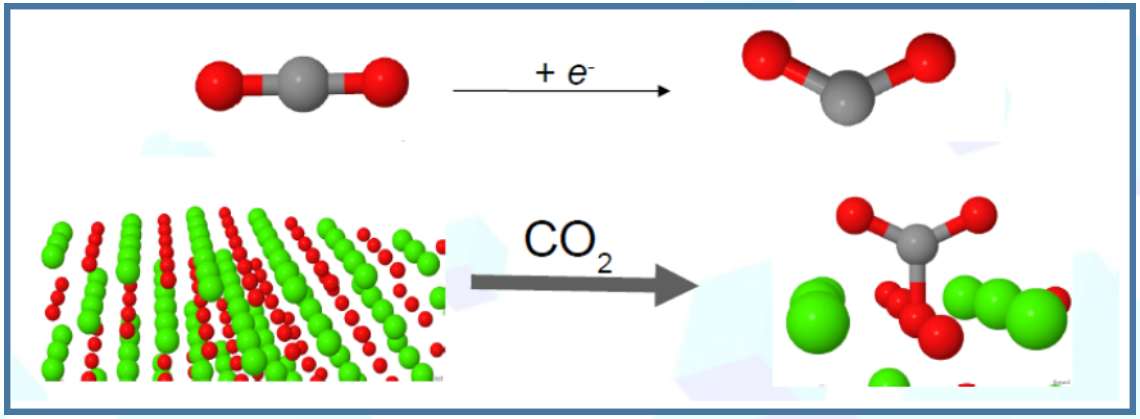
Turning Greenhouse Gases into Useful Chemicals and Fuels



We need an efficient catalyst!

Turning Greenhouse Gases into Useful Chemicals and Fuels

NOVEL MATERIALS DISCOVERY



Prediction of new metal-oxide catalysts for CO_2 reduction.

A combination of adsorption and distortion of the molecule.

Training data: ~200 Me-O surfaces × adsorption-sites, DFT CO_2 adsorption
Reference data: experimental data on catalytic activity.

Investigation of adsorption energy, OCO angle reduction, CO bond elongation, and more as activation indicators on the basis of features coming from isolated Me atoms, bulk properties, and pristine surfaces.

The study lead to a proposal of a definition of CO_2 activation.

Turning Greenhouse Gases into Useful Chemicals and Fuels

NOVEL MATERIALS DISCOVERY

Oxides:

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

A^{2+}: Mg, Ca, Sr, Ba
A^{3+}(**B**$^{3+}$): Al, Ga, In, Sc, Y, La
B^{4+}: Ti, Zr, Si, Ge, Sn
A^{+}: Li, Na, K, Rb, Cs;
B^{5+}: Nb, V, Sb

Consider surfaces of many different materials and all possibly relevant surface sites: Which materials (and surface sites) are catalytically active?

Global vs local learning: subgroup discovery

NOVEL MATERIALS DISCOVERY

Atom properties
Bulk properties
Pristine surface properties
CO_2 properties
Candidate descriptors

Global vs local learning: subgroup discovery

NOVEL MATERIALS DISCOVERY

Atom properties
Bulk properties
Pristine surface properties
CO_2 properties
Candidate descriptors

SGD

Target properties
CO2 on surface properties:
- E_{ads}
- bending angle
- CO bond length
Candidate Indicators of activation

Global vs local learning: subgroup discovery

NOVEL MATERIALS DISCOVERY

Atom properties
Bulk properties
Pristine surface properties
CO_2 properties
Candidate descriptors

SGD

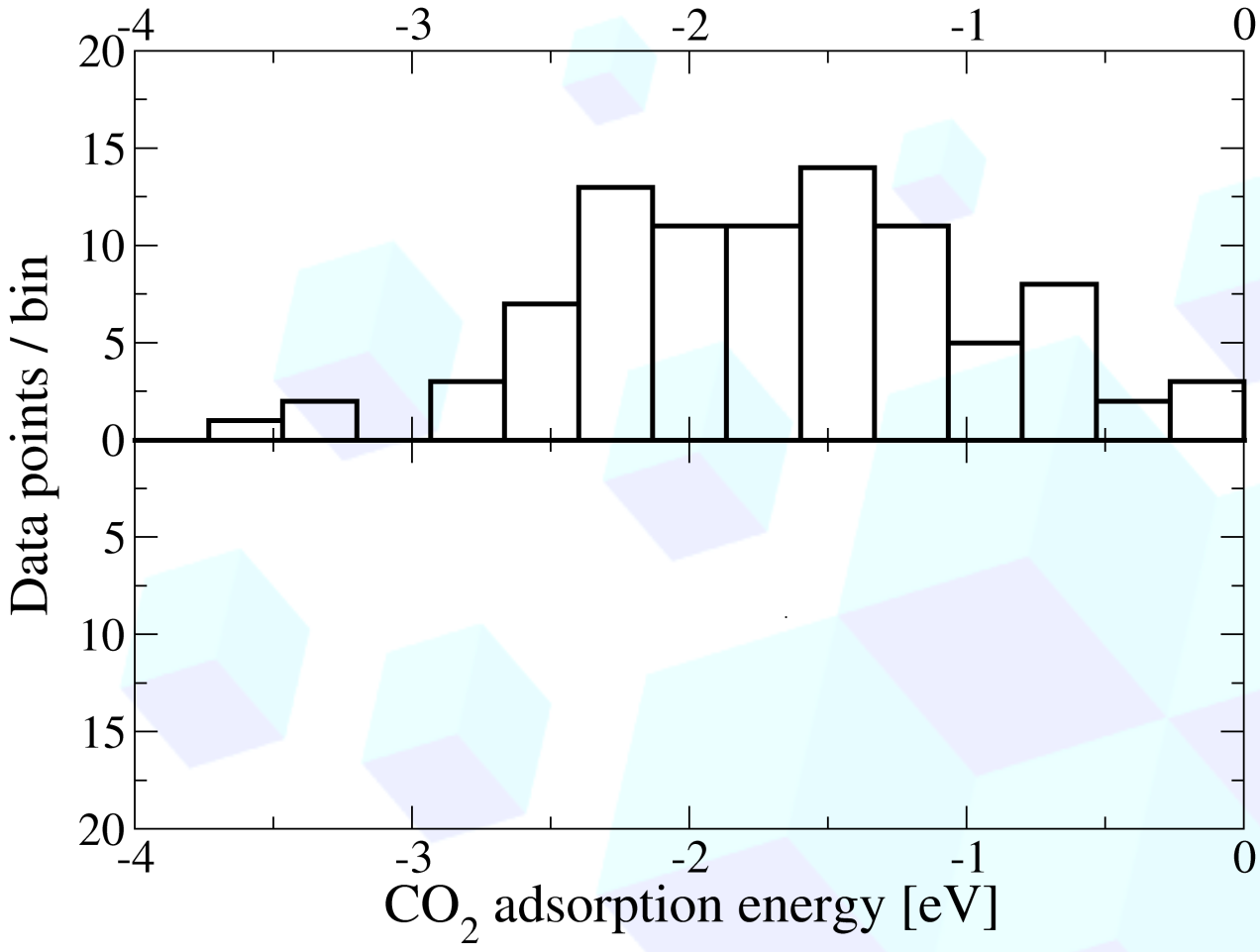
Target properties
CO2 on surface properties:
- E_{ads}
- bending angle
- CO bond length

Candidate Indicators of activation

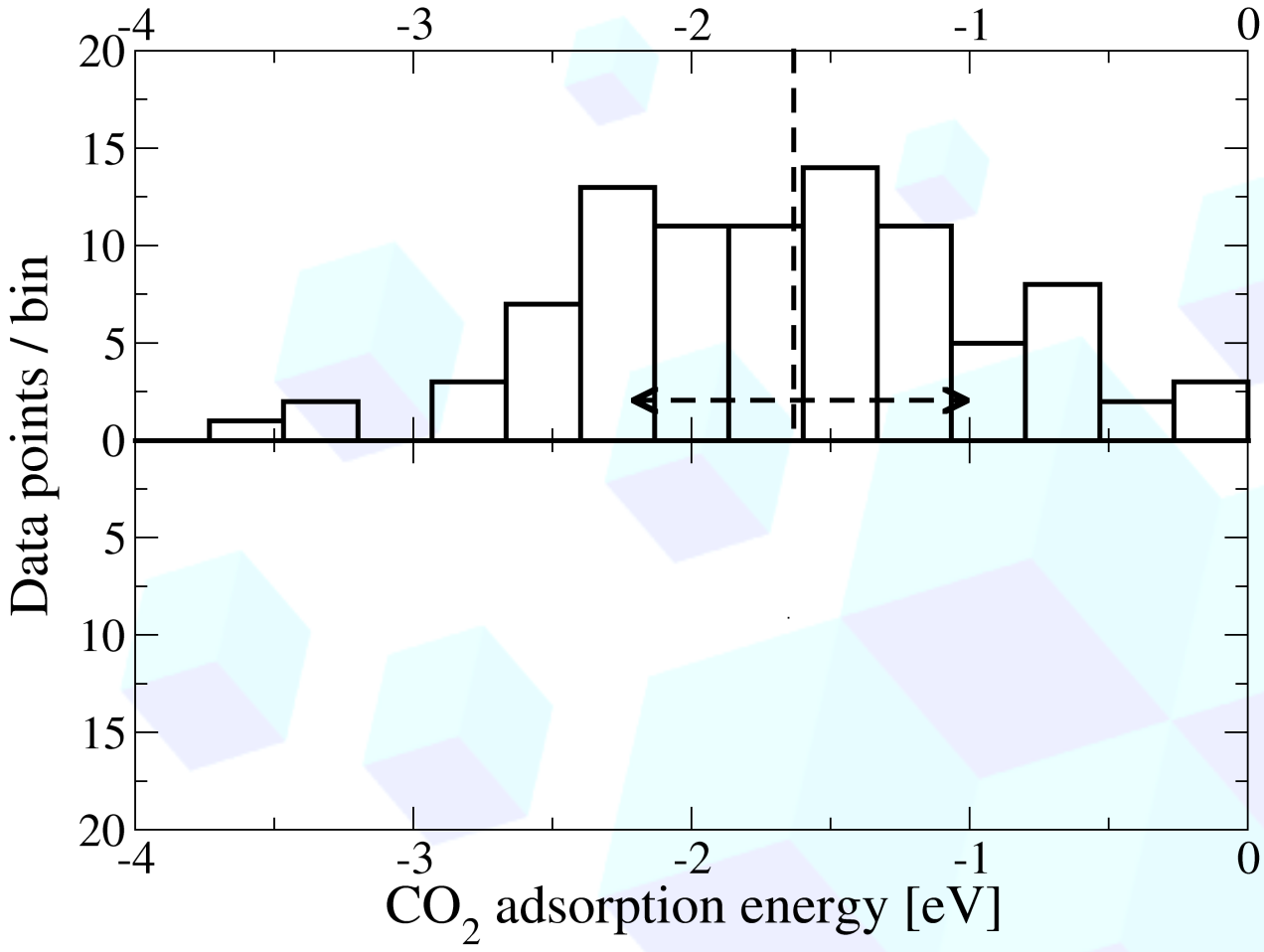
Unsupervised learning

Classification
Catalytically active vs inactive

Subgroup discovery by example



Subgroup discovery by example



Subgroup discovery by example

$$\underset{SG \subset P}{\operatorname{argmax}}\, U = \frac{\#SG}{\#P}\left(1 - \frac{\operatorname{mad}(SG)}{\operatorname{mad}(P)}\right)\left|\operatorname{med}(SG) - \operatorname{med}(P)\right|$$

Boley, Goldsmith, LMG & Vreeken, Data Min. Knowl. Disc. **31**, 1391 (2017)
Goldsmith, Boley, Vreeken, Scheffler & LMG, New J. Phys. **19**, 013031 (2017)

Subgroup discovery by example

Size of subgroup SG

$$\underset{SG \subset P}{\operatorname{argmax}}\, U = \frac{\#SG}{\#P}\left(1 - \frac{\operatorname{mad}(SG)}{\operatorname{mad}(P)}\right)\left|\operatorname{med}(SG) - \operatorname{med}(P)\right|$$

Size of full set P

Boley, Goldsmith, LMG & Vreeken, Data Min. Knowl. Disc. **31**, 1391 (2017)
Goldsmith, Boley, Vreeken, Scheffler & LMG, New J. Phys. **19**, 013031 (2017)

Subgroup discovery by example

Size of subgroup SG

Mean absolute deviation from the median (spread of distribution)

$$\underset{SG \subset P}{\operatorname{argmax}}\, U = \frac{\#SG}{\#P}\left(1 - \frac{\operatorname{mad}(SG)}{\operatorname{mad}(P)}\right)|\operatorname{med}(SG) - \operatorname{med}(P)|$$

Size of full set P

Boley, Goldsmith, LMG & Vreeken, Data Min. Knowl. Disc. **31**, 1391 (2017)
Goldsmith, Boley, Vreeken, Scheffler & LMG, New J. Phys. **19**, 013031 (2017)

Subgroup discovery by example

Size of subgroup SG

Mean absolute deviation from the median (spread of distribution)

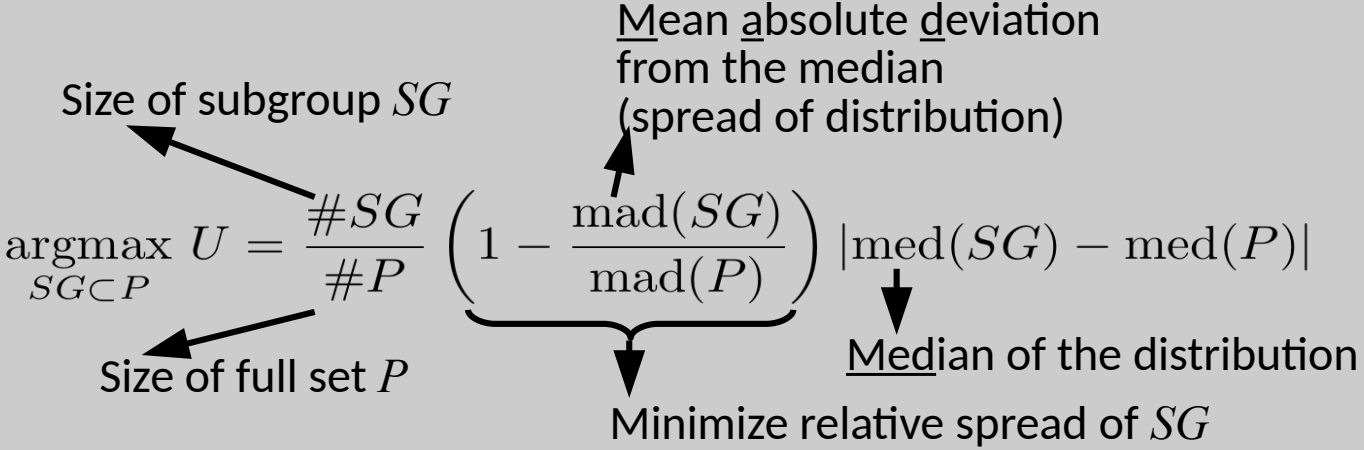
$$\underset{SG \subset P}{\operatorname{argmax}}\, U = \frac{\#SG}{\#P}\left(1 - \frac{\operatorname{mad}(SG)}{\operatorname{mad}(P)}\right)|\operatorname{med}(SG) - \operatorname{med}(P)|$$

Size of full set P

Minimize relative spread of SG

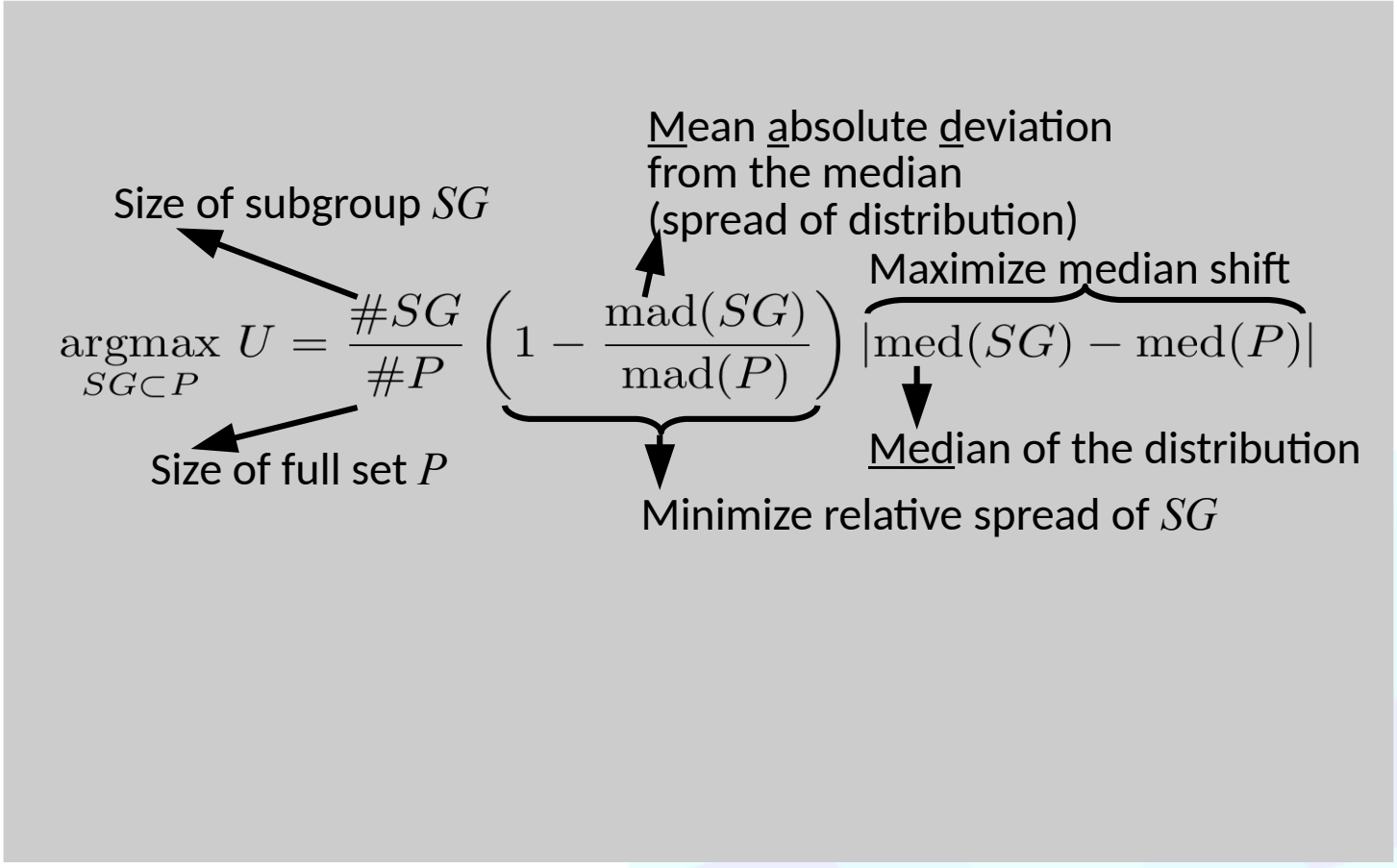
Boley, Goldsmith, LMG & Vreeken, Data Min. Knowl. Disc. **31**, 1391 (2017)
Goldsmith, Boley, Vreeken, Scheffler & LMG, New J. Phys. **19**, 013031 (2017)

Subgroup discovery by example



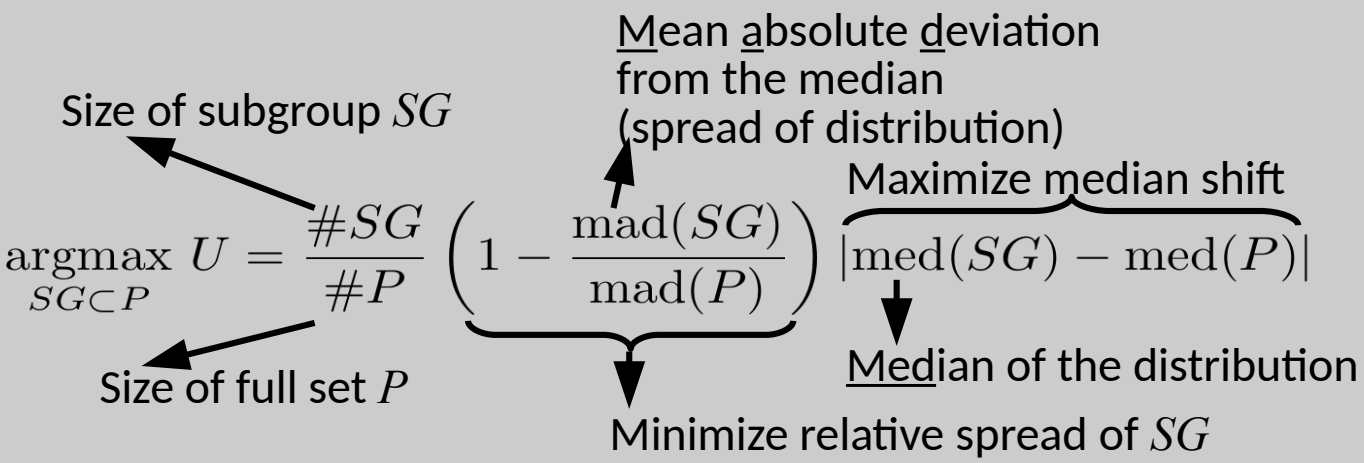
Boley, Goldsmith, LMG & Vreeken, Data Min. Knowl. Disc. **31**, 1391 (2017)
Goldsmith, Boley, Vreeken, Scheffler & LMG, New J. Phys. **19**, 013031 (2017)

Subgroup discovery by example



Boley, Goldsmith, LMG & Vreeken, Data Min. Knowl. Disc. **31**, 1391 (2017)
Goldsmith, Boley, Vreeken, Scheffler & LMG, New J. Phys. **19**, 013031 (2017)

Subgroup discovery by example

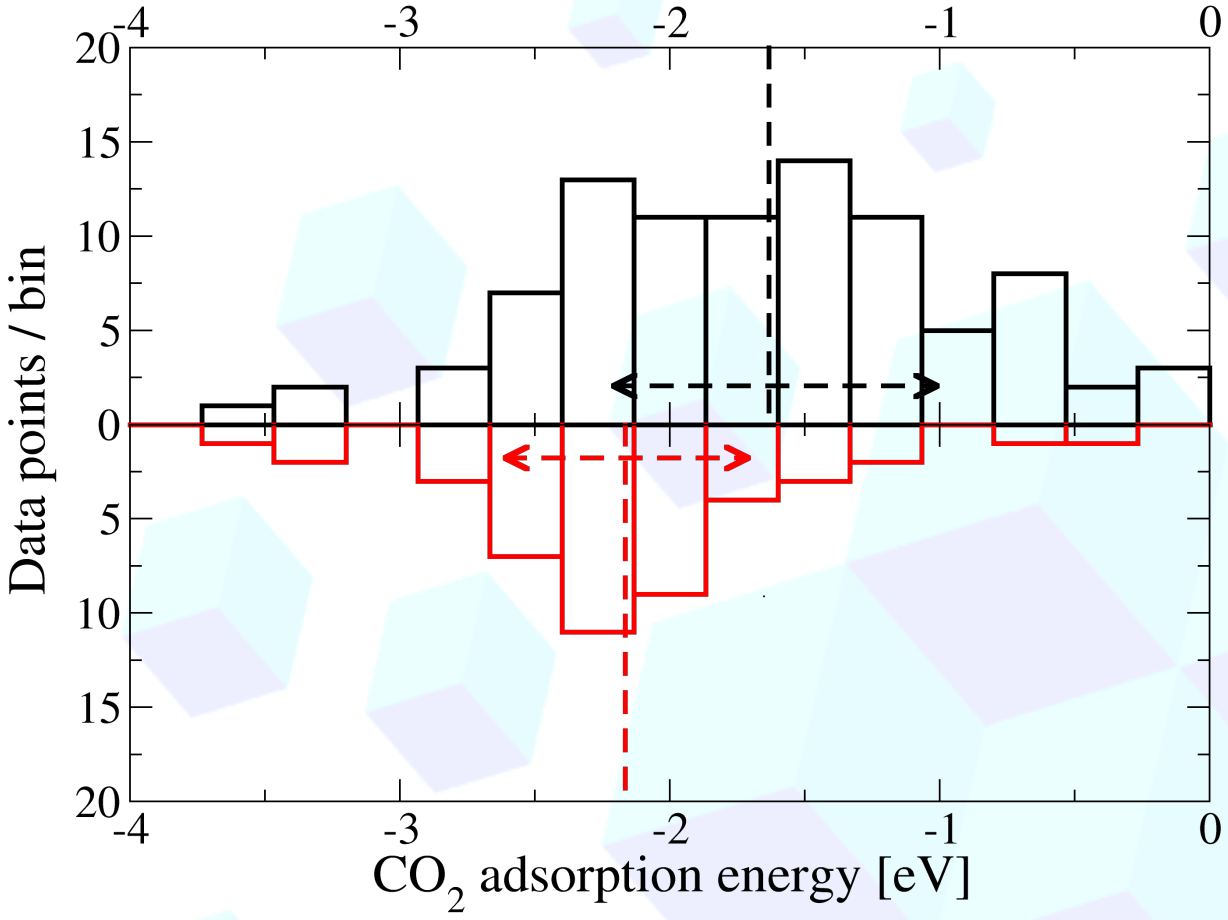


SG is *described* by a *selector*,
a conjunction of statements ($s_1 \wedge s_2 \wedge \ldots$)
about a list of given features e.g.,
s_1 = surface energy larger than … ,
s_2 = center of surface-O projected *p*-band less than …

Boley, Goldsmith, LMG & Vreeken, Data Min. Knowl. Disc. **31**, 1391 (2017)
Goldsmith, Boley, Vreeken, Scheffler & LMG, New J. Phys. **19**, 013031 (2017)

Subgroup discovery by example

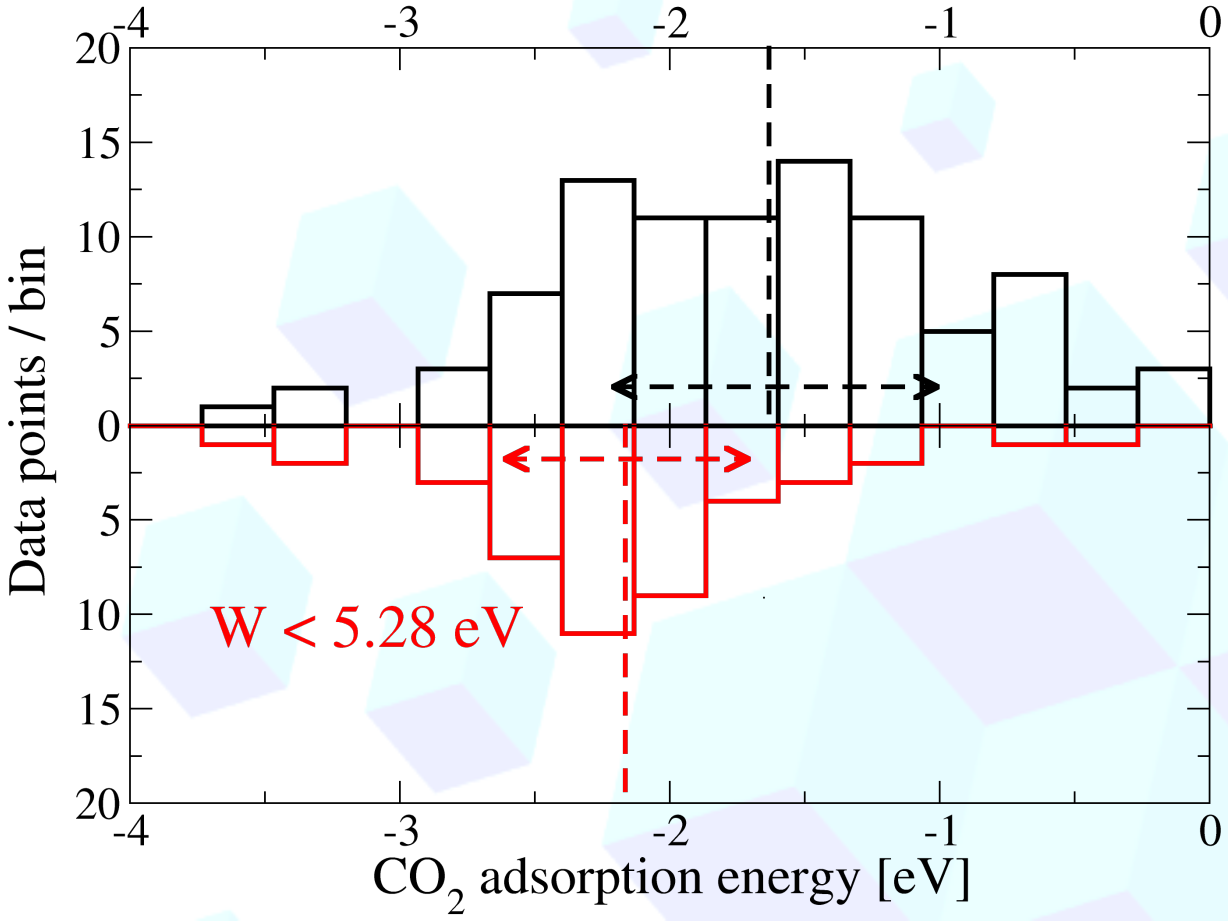
$$\underset{SG \subset P}{\operatorname{argmax}}\, U = \frac{\#SG}{\#P}\left(1 - \frac{\mathrm{mad}(SG)}{\mathrm{mad}(P)}\right)\left|\mathrm{med}(SG) - \mathrm{med}(P)\right|$$



Mazheika, Wang, LMG, Illas, Scheffler, Levchenko, in preparation

Subgroup discovery by example

$$\underset{SG \subset P}{\operatorname{argmax}}\, U = \frac{\#SG}{\#P}\left(1 - \frac{\operatorname{mad}(SG)}{\operatorname{mad}(P)}\right)\left|\operatorname{med}(SG) - \operatorname{med}(P)\right|$$

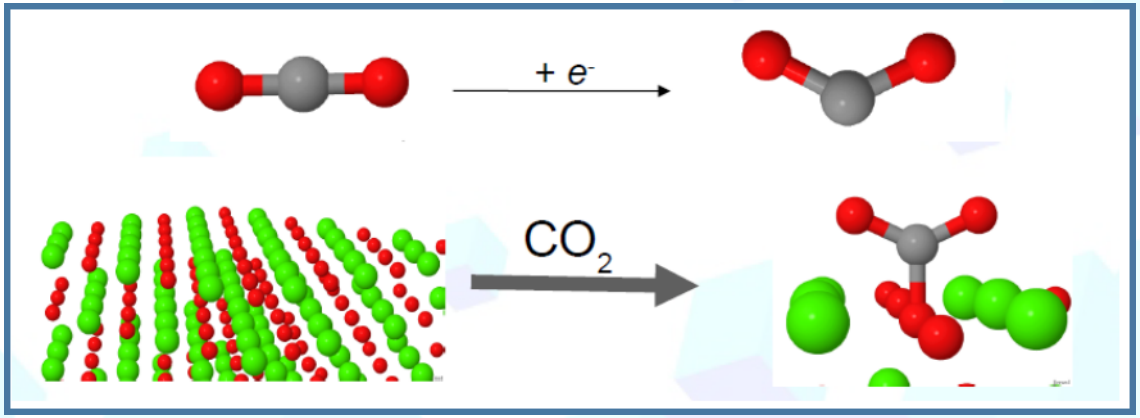


Mazheika, Wang, LMG, Illas, Scheffler, Levchenko, in preparation

Global vs local learning: subgroup discovery

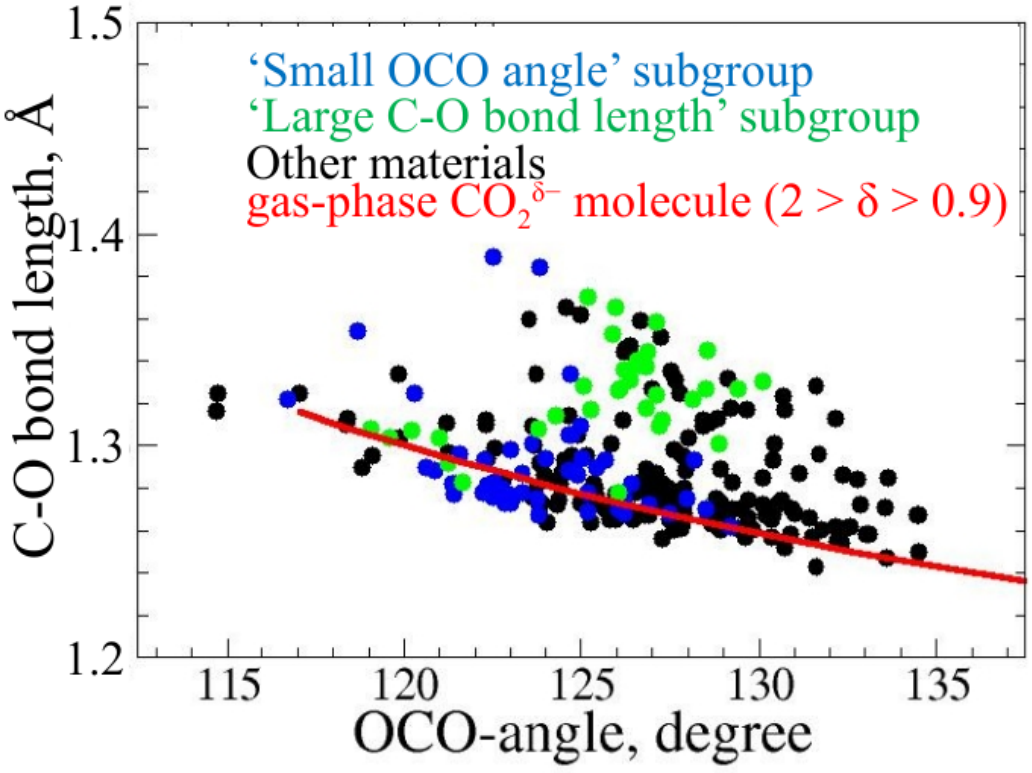
Subgroup identification:

- Define a '**target property**'
  O-C-O angle
  C-O bond length



- Simultaneously:
  - Minimize the width of the target-property distribution.
  - Maximize the distance between the median of the target-property distribution and that of the whole data set.
  - Maximize the size of the subgroup.

Turning Greenhouse Gases into Useful Chemicals and Fuels

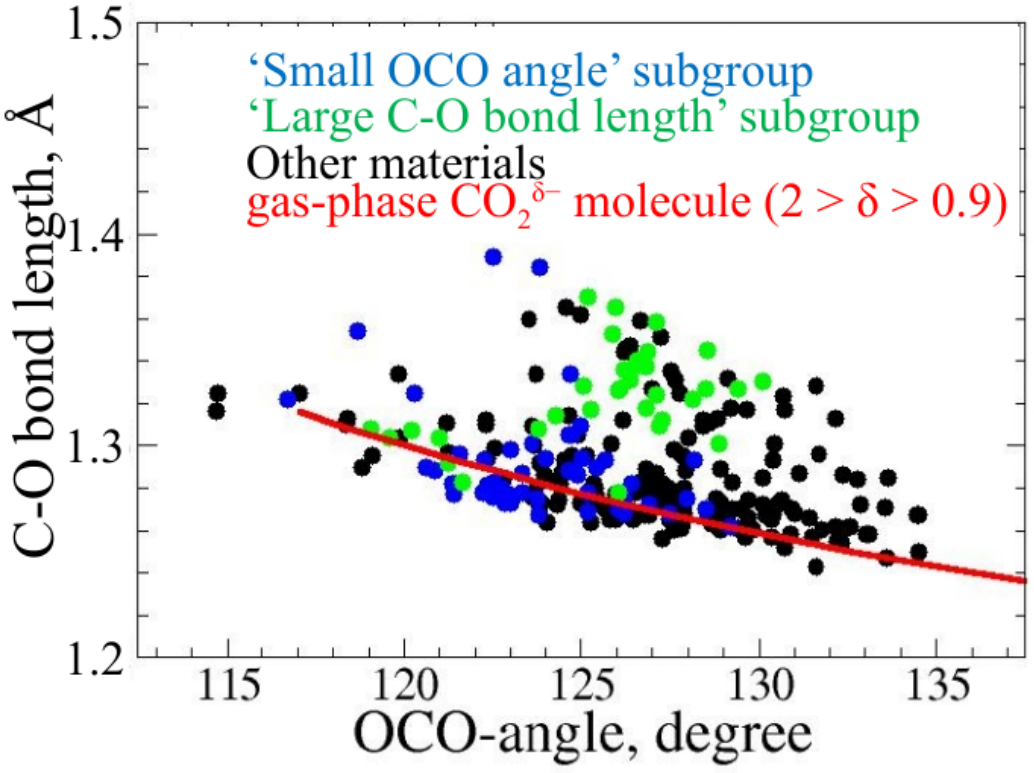


Turning Greenhouse Gases into Useful Chemicals and Fuels

$VBM < -5.14$ eV
(wrt vacuum)

Min. of Hirschfeld charges of the A and B atoms
$q_{min} < 0.48$ e$^-$

Distance between the O surface atom and its second-nearest neighbor cation
$d_2 > 2.26$ Å

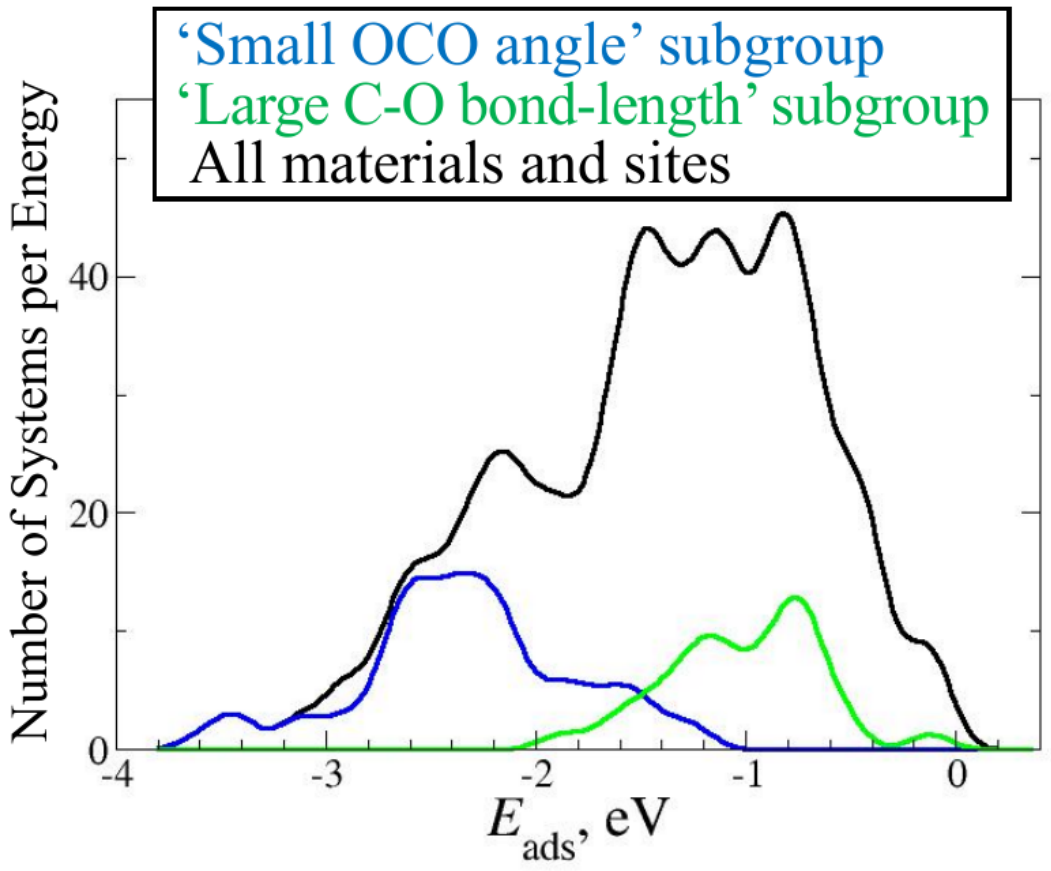


Max. of O 2*p* DOS
$M > -6.0$ eV

Distance between O surface atom and its nearest neighbor cation $d_1 > 1.8$ Å

Distance between the O surface atom and its second-nearest neighbor cation
$d_2 > 2.12$ Å

Turning Greenhouse Gases into Useful Chemicals and Fuels



Most known materials with good catalytic performance belong to the 'large C-O bond length' subgroup.

From the "bad-performance materials", none belongs to the green subgroup.

Global vs local learning: subgroup discovery

NOVEL MATERIALS DISCOVERY

Atom properties
Bulk properties
Pristine surface properties
CO2 properties
Candidate descriptors

SGD

Target properties
CO2 on surface properties:
- Eads
- bending angle
- CO bond length

Candidate Indicators of activation

Next:
Predict ‘C-O bond length’ from descriptor:
- atom properties
- bulk properties
- ideal (geometrical) surface properties

Unsupervised learning

Classification
Catalytically active vs inactive

Global vs local learning: subgroup discovery

Subgroup identification:

- Define a '**target property**'
  Mean Absolute Error of the *predicted* cohesive energy

Global vs local learning: subgroup discovery

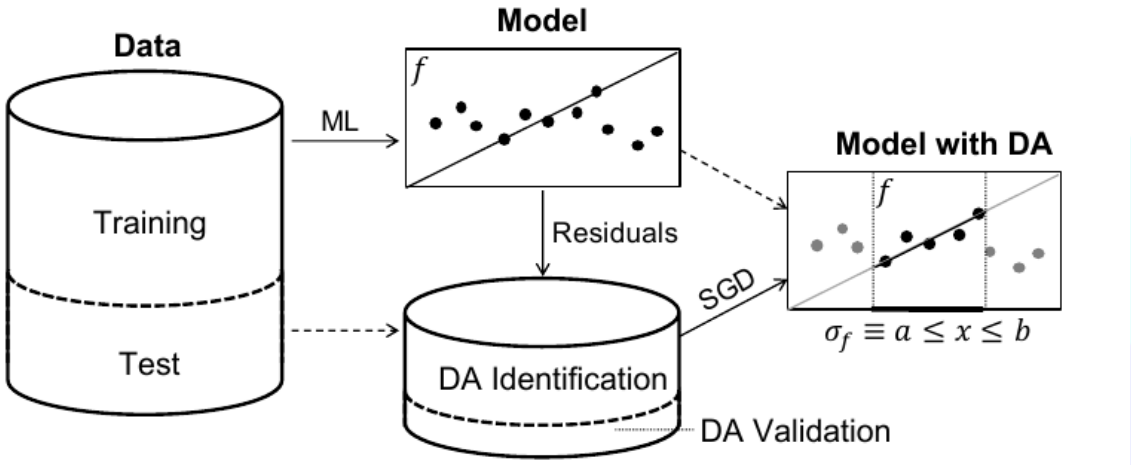
***Subgroup identification*:**

- Define a '**target property**'
  Mean Absolute Error of the *predicted* cohesive energy
- Simultaneously:
  - Minimize the width of the target-property distribution.
  - Maximize the distance between the median of the target-property distribution and that of the whole data set.
  - Maximize the size of the subgroup.

NOMAD

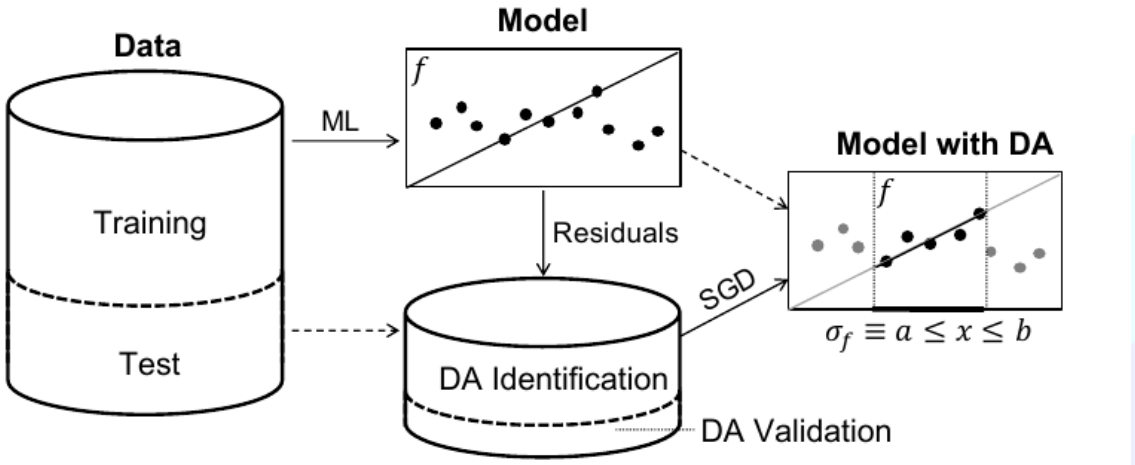
NOVEL MATERIALS DISCOVERY

Domain of applicability (DA) of ML models



Domain of applicability (DA) of ML models

NOVEL MATERIALS DISCOVERY



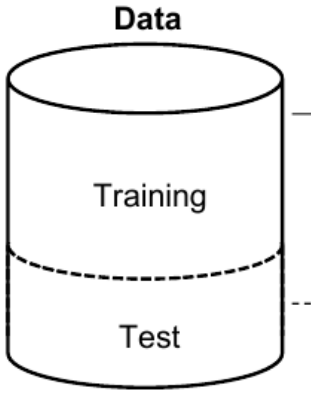
Example: Data from NOMAD-Kaggle-2018

Competition on transparent, conducting oxides: $(Al_xGa_yIn_z)_2O_3$ (for 6 space groups and up to 80 atoms/unit cell).

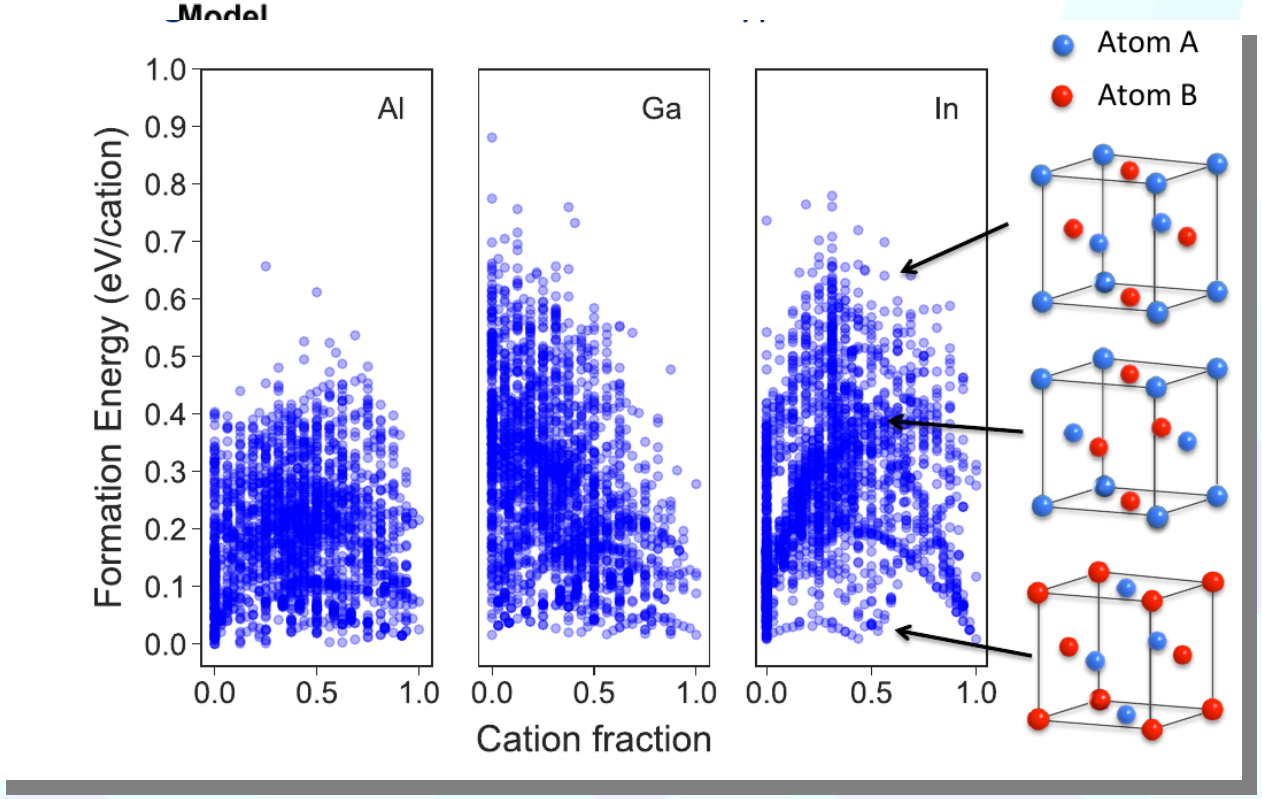
Sutton et al., NPJ Comp. Materials (2019)

Domain of applicability (DA) of ML models

NOVEL MATERIALS DISCOVERY



Example: Data from NOMAD-Kag
Competition on transparent, cond
up to 80 atoms/unit cell).



Sutton et al., NPJ Comp. Materials (2019)

Domain of applicability (DA) of ML models

Example: $(Al_xGa_yIn_z)_2O_3$

with Gaussian-kernel KRR and different representations

Common representation for SGD: lattice-vector lengths and angles, volume per atom, # atoms/unit cell, composition (%), average *nn* distances (Al-Al, Al-Ga, …)

ML model	all data (meV/cation)	DA (meV/cation)	selectors defining the DA
n-gram	15.2		
SOAP	14.5		
MBTR	13.9		

Mean Absolute Error of the predicted cohesive energy

Domain of applicability (DA) of ML models

NOVEL MATERIALS DISCOVERY

Feature type	Feature label	Feature definition (units)
Unit cell	a, b, c	Lattice-vector lengths sorted from largest (a) to smallest (c) (Å)
	α	angle between b and c (°)
	β	angle between a and c (°)
	γ	angle between a and b (°)
	$\frac{V}{V_{atom}}$	volume of unit cell divided by atomic volumes derived from covalent radii
	N	number of atoms
Composition	%Al, %Ga, %In	number of cations divided by total number of cations
Compositionally averaged * atomic properties	E_g	PBE band gap energy
Structural	$R_{\{Al,Ga,In,O\}-\{Al,Ga,In,O\}}$	average nearest-neighbor distance between, e.g., Al, Ga, In, and oxygen, within the first coordination shell

Domain of applicability (DA) of ML models

NOVEL MATERIALS DISCOVERY

Example: $(Al_xGa_yIn_z)_2O_3$

with Gaussian-kernel KRR and different representations

Common representation for SGD: lattice-vector lengths and angles, volume per atom, # atoms/unit cell, composition (%), average *nn* distances (Al-Al, Al-Ga, ...)

ML model	all data (meV/cation)	DA (meV/cation)	selectors defining the DA
n-gram	15.2	11.41	$b \geq 5.59$ Å $\wedge$ $\gamma < 90.35°$ $\wedge$ $R_{\text{Al–O}} \leq 2.06$Å $\wedge$ $R_{\text{Ga–O}} \leq 2.07$Å
SOAP	14.5	11.25	$a/c \leq 3.89$ $\wedge$ $\gamma < 90.35°$ $\wedge$ $\beta \geq 88.68°$
MBTR	13.9	8.03	$N \geq 50$ $\wedge$ $\gamma < 90.35°$ $\wedge$ $R_{\text{Al-O}} \leq 2.06$ Å

Mean Absolute Error of the predicted cohesive energy

Acknowledgements

Subgroup Discovery for Domain of Applicability: Mario Boley, Christopher Sutton, Matthias Rupp, Jilles Vreeken

Subgroup discovery for CO_2 activation: Aliaksei Mazheika, Sergey Levchenko, Yanggang Wang, Rosendo Valero, Francesc Illas

And: Matthias Scheffler

NOMAD has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 676580.