Uncovering structure-property relationships by applying subgroup discovery to materials-science data

Bryan R. Goldsmith, Mario Boley, Jilles Vreeken,
Luca M. Ghiringhelli and Matthias Scheffler

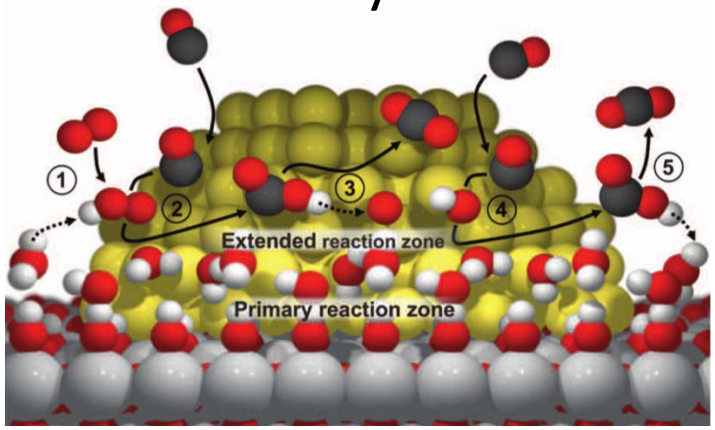
Fritz Haber Institute of the Max Planck Society
Theory Department

NOMAD
Novel Materials Discovery Laboratory
- A European Center of Excellence

Alexander von Humboldt
Stiftung/Foundation

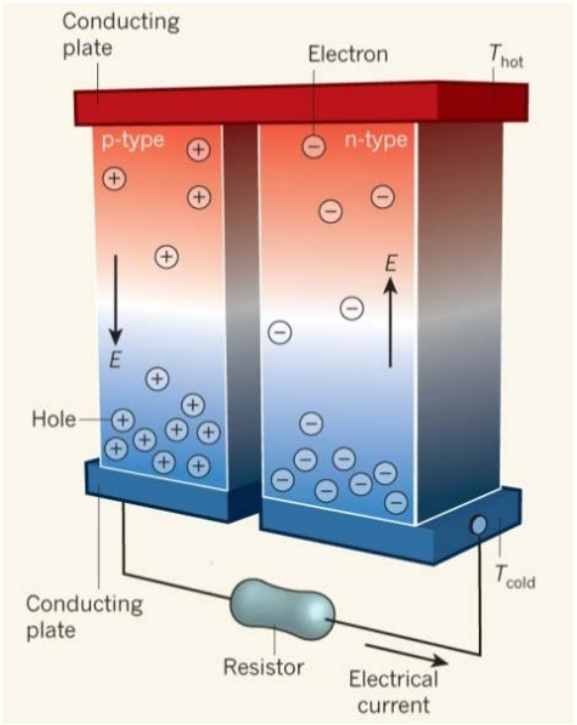
Designing materials requires an understanding of the mechanisms underlying a materials function

Catalysts

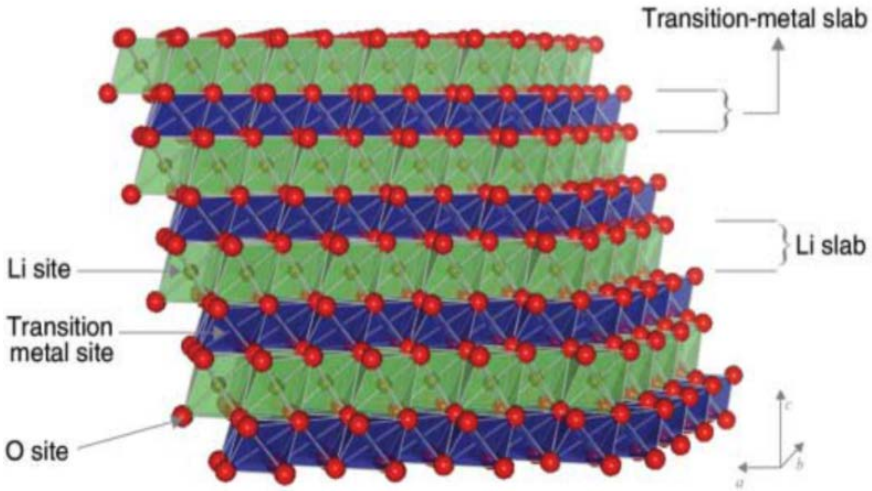


[1]

Thermoelectrics



Batteries



[2]

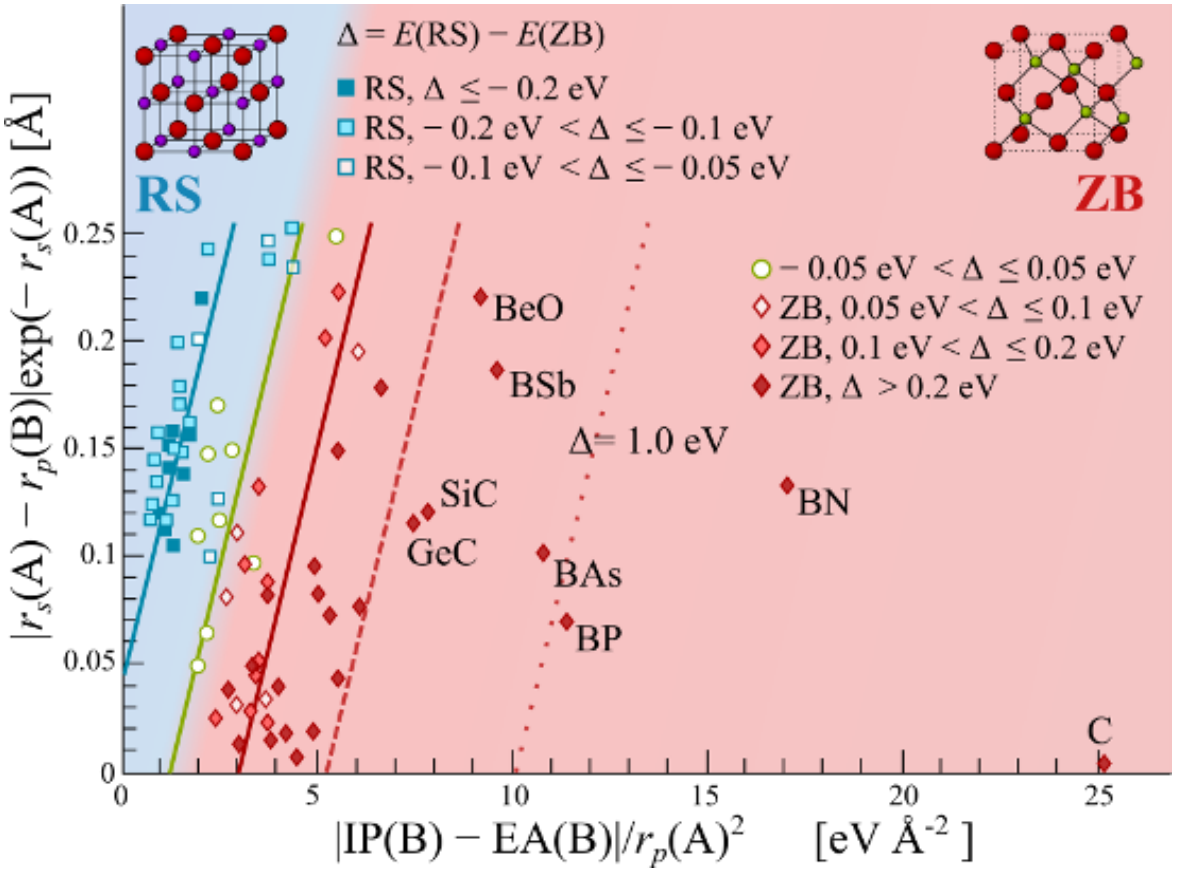
[1] J. Saavedra *et al. Science* 345, 1599 (2014)

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

Data analytics can help identify patterns and descriptors in big-data of materials

Descriptor → Property

Descriptor = function(atomic or material features)



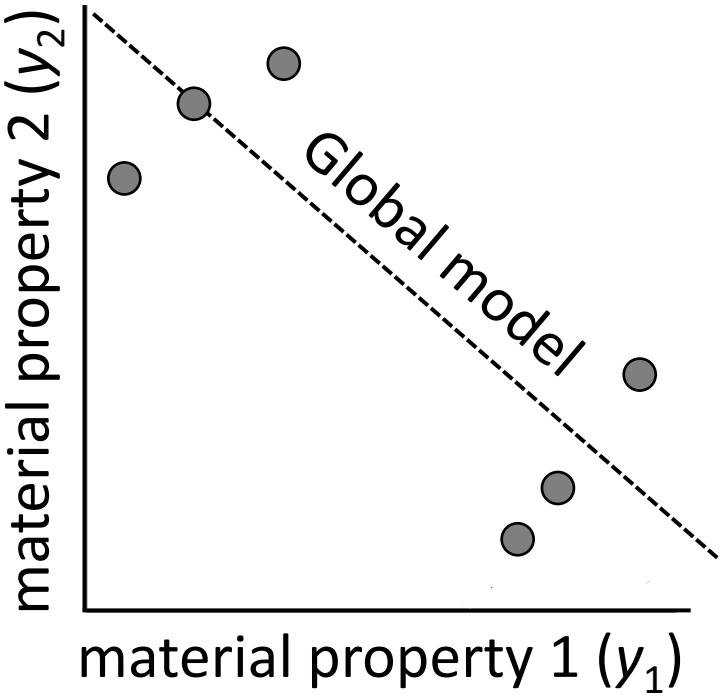
Feature selection via **LASSO+l_0** [1]

[1] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, *PRL 114*, 105503 (2015)

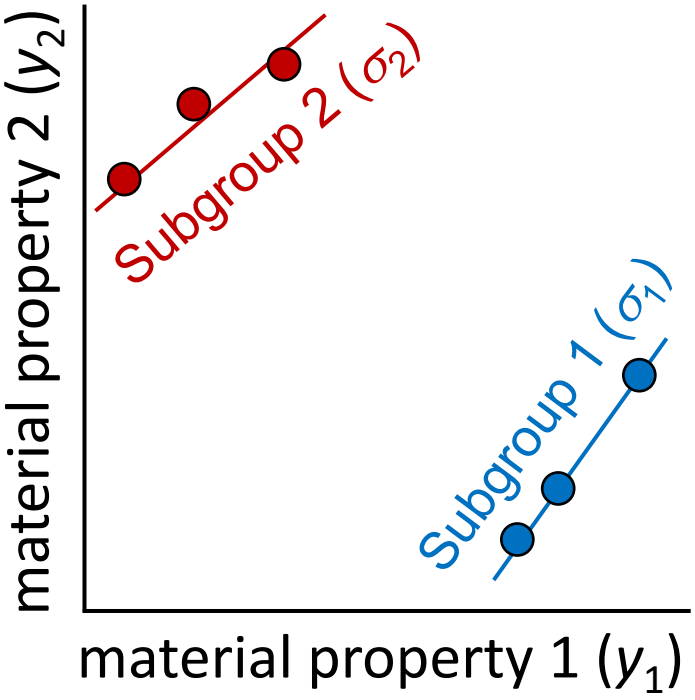
Typically one focuses on the inference of a global prediction model for some property of interest (e.g., LASSO, KRR, l_0)

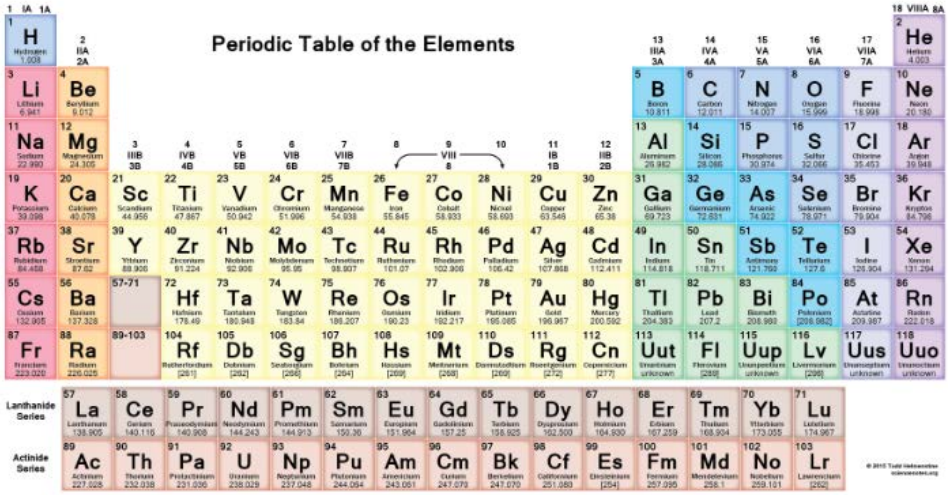
Underlying mechanisms can change across materials

Relations between subsets of data may be important



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

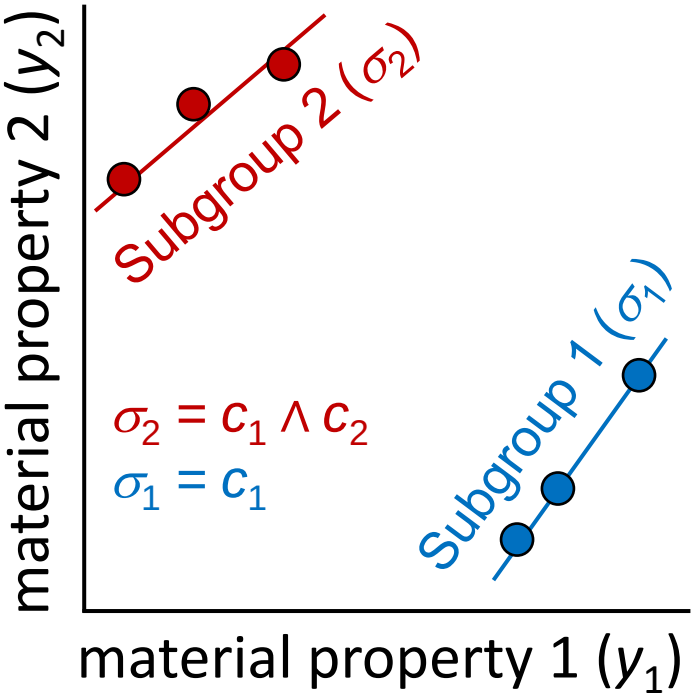




Transition metals are a subgroup
Halogens are a subgroup
etc…

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



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

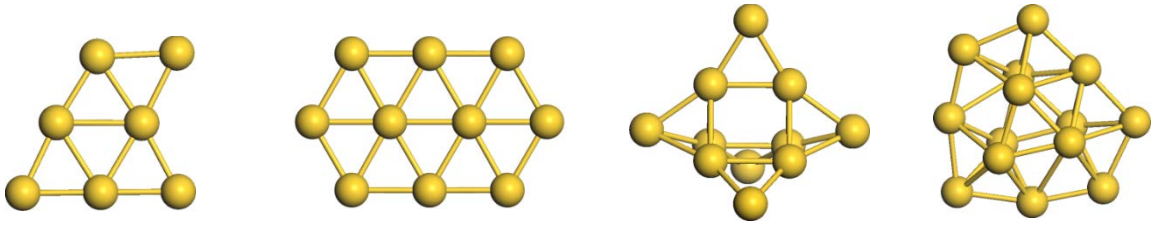
$\frac{|\mathrm{ext}(\sigma)|}{|P|}$ is the coverage of points where σ is true

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

Two applications of subgroup discovery are presented here

1. Gas Phase Gold Clusters (of sizes 5-14 atoms)

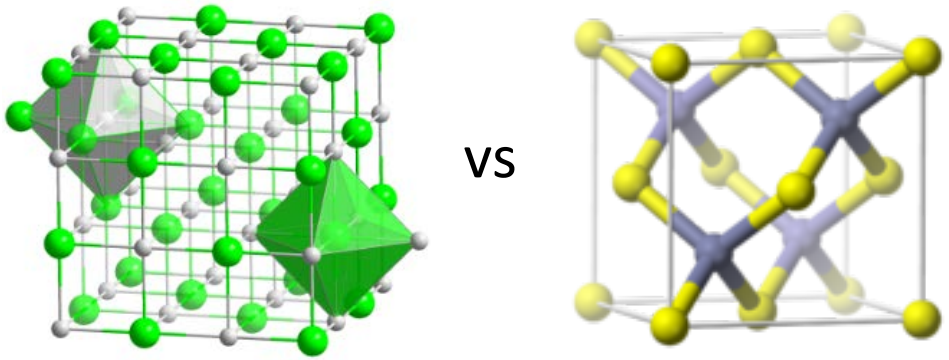
Display interesting optical, chemical, and electronic properties



24 400 gold cluster configurations in total

2. Classification of 82 Octet Binary Semiconductors

Rocksalt (RS) vs Zincblende (ZB)



Rediscover simple insight about HOMO-LUMO gap

24 400 gold cluster configurations (of sizes 5-14) in the gas phase

Choose **target property**

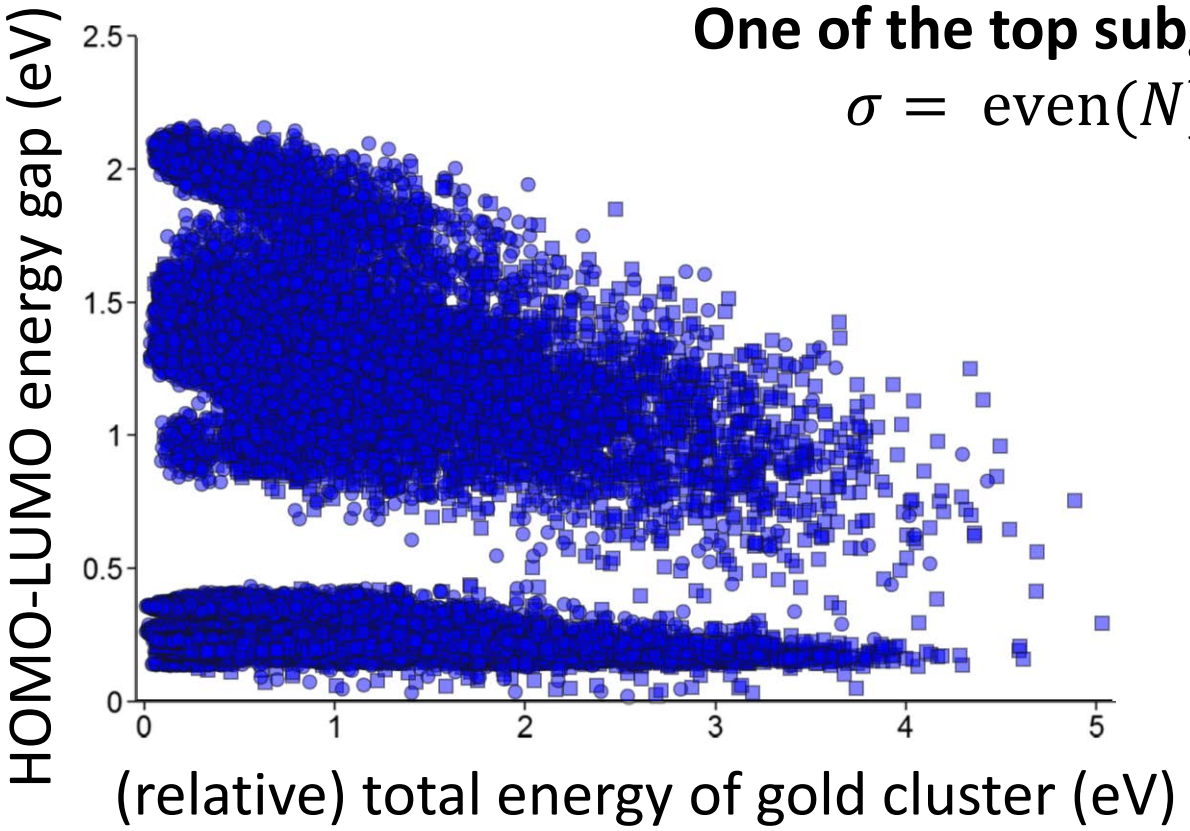
HOMO-LUMO energy gap

Choose **variation reduction utility function**

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

One of the top subgroup selectors found

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



Rediscover simple insight about HOMO-LUMO gap

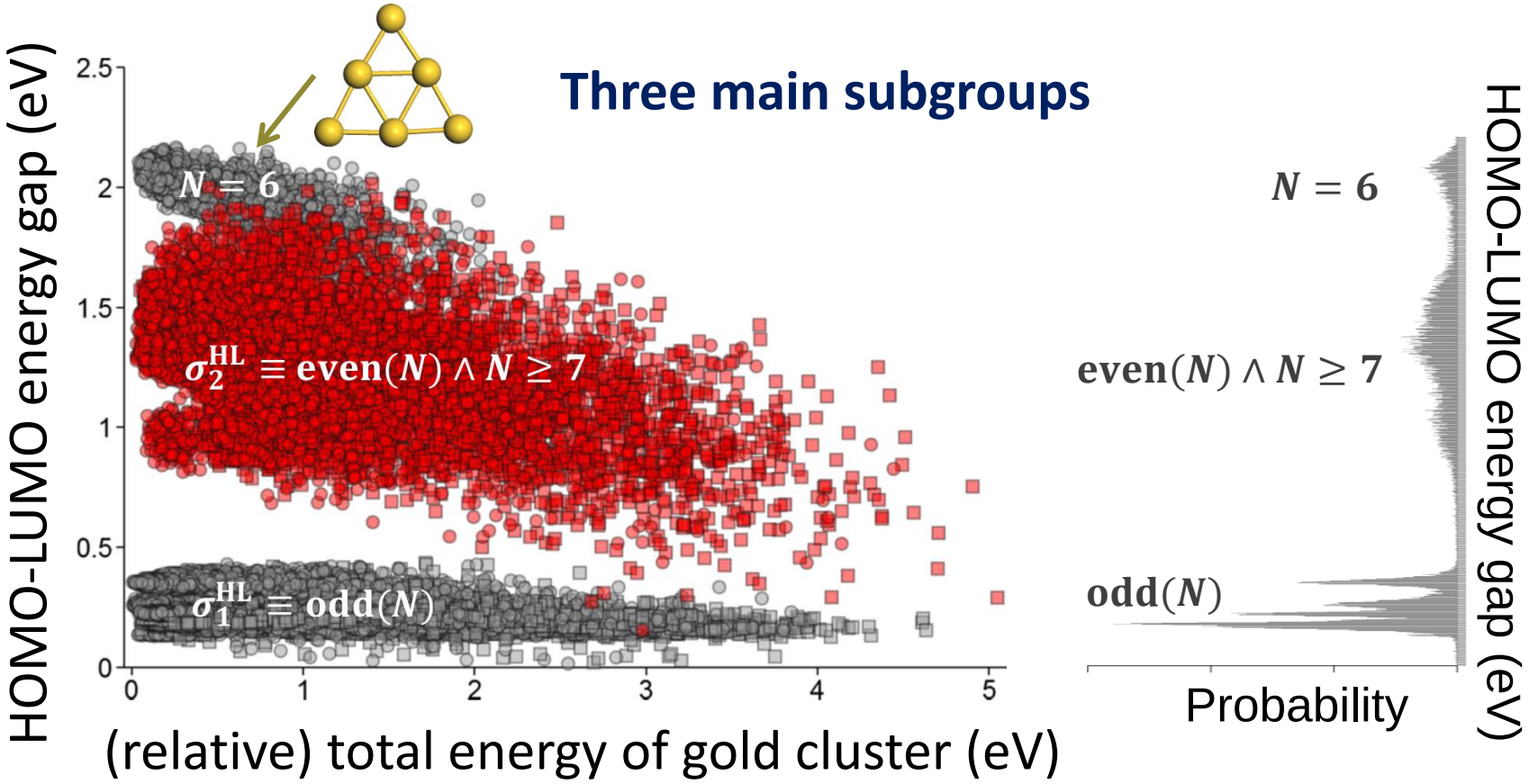
24 400 gold cluster configurations (of sizes 5-14) in the gas phase

Choose **target property**

HOMO-LUMO energy gap

Choose **variation reduction utility function**

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

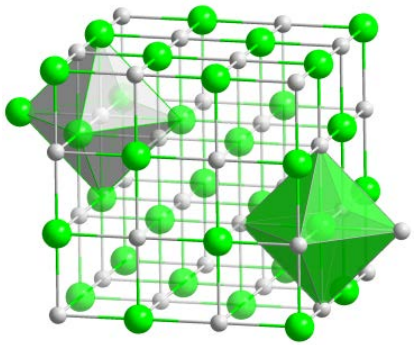


Can subgroup discovery find local models for the 82 octet binaries that describe zincblende and rocksalt?

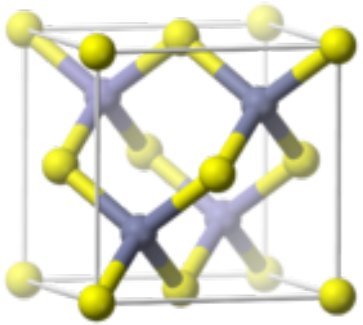
Target = sign ($E_{\text{rocksalt}} - E_{\text{zincblende}}$)

Subgroup 1 → Rocksalt (RS)

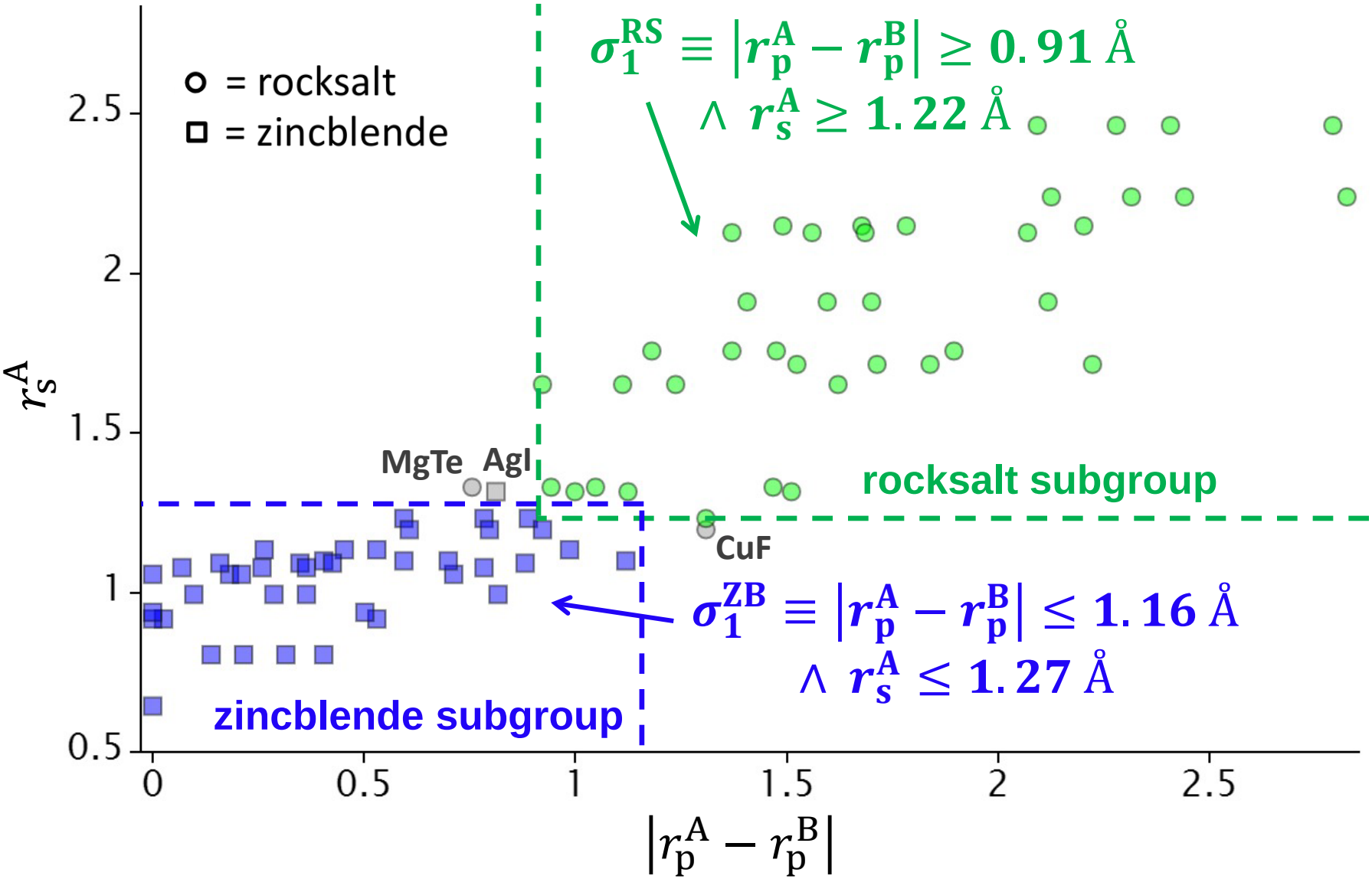
Subgroup 2 → Zincblende (ZB)



vs



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



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

Subgroup discovery can help identify
physically meaningful models in materials-science data

Other analytics tools being developed to find descriptors
e.g., LASSO+l_0 and sure independent screening
(Runhai Ouyang, Luca M. Ghiringhelli)

Materials uploaded on NOMAD repository
http://nomad-repository.eu/cms/

Big-data analytics toolkit and tutorials (including subgroup discovery)
https://www.nomad-coe.eu/

- Bigger data and harder problems

Acknowledgements

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

max planck institut
informatik