

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



*Novel Materials
Discovery Laboratory*

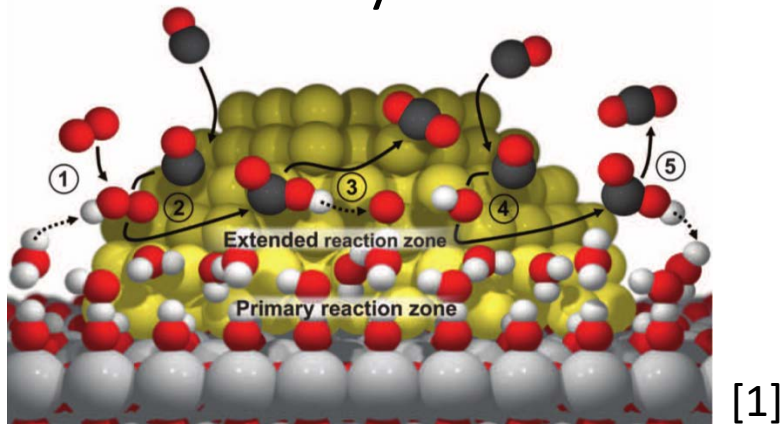
- A European Center of Excellence



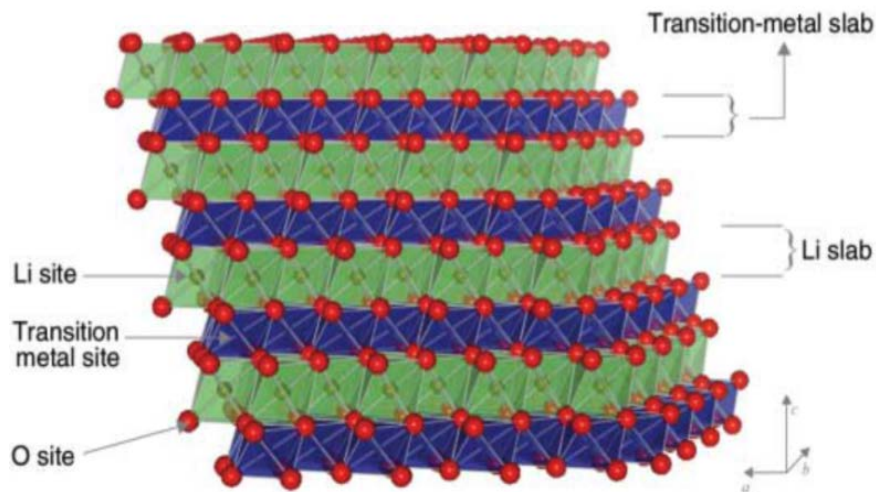
Alexander von Humboldt
Stiftung / Foundation

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

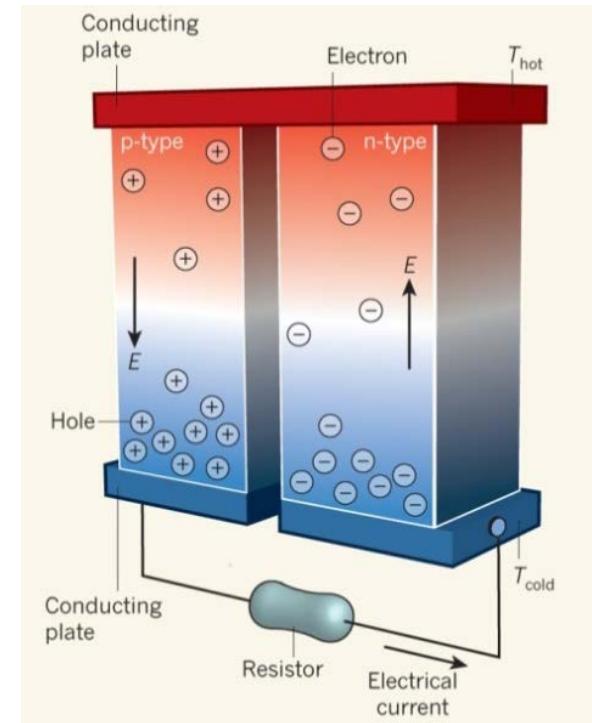
Catalysts



Batteries



Thermoelectrics



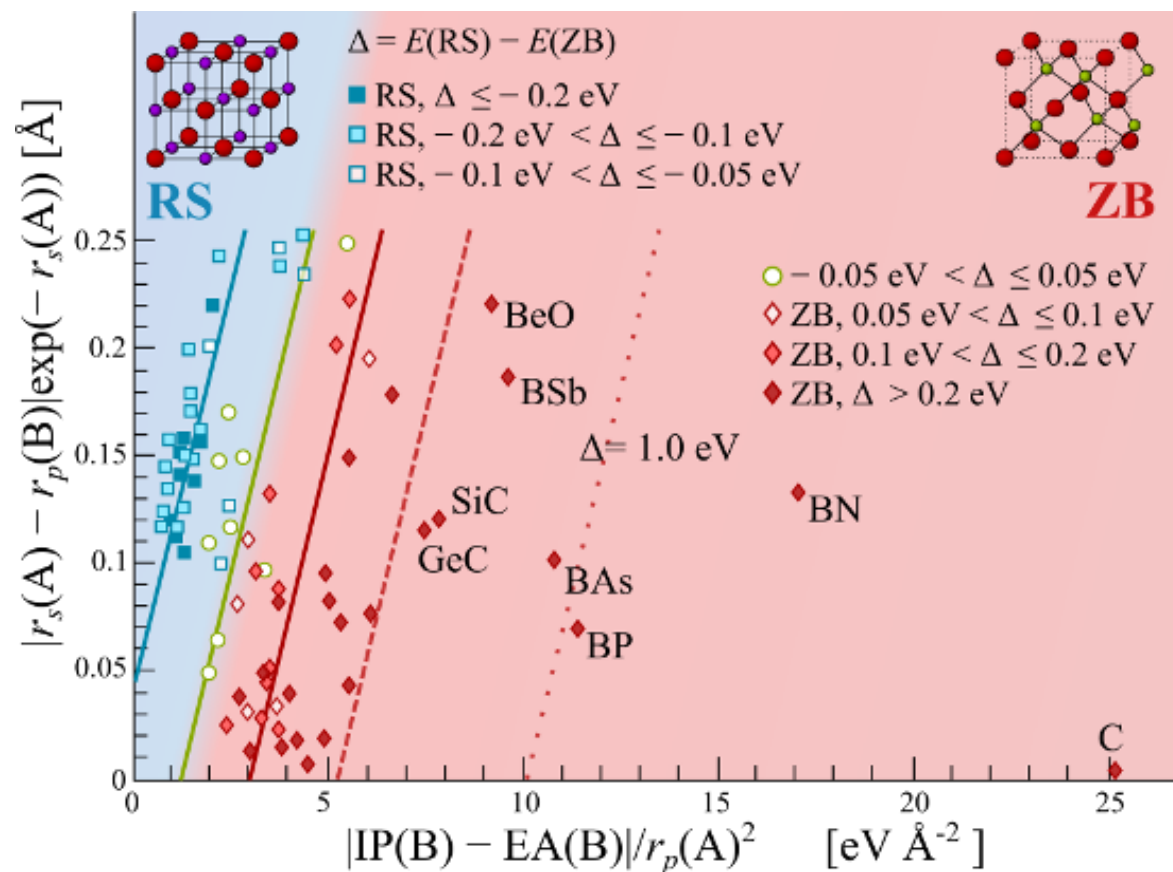
[1] J. Saavedra *et al.* *Science* 345, 1599 (2014)

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

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

Descriptor \rightarrow Property

Descriptor = function(atomic or material features)

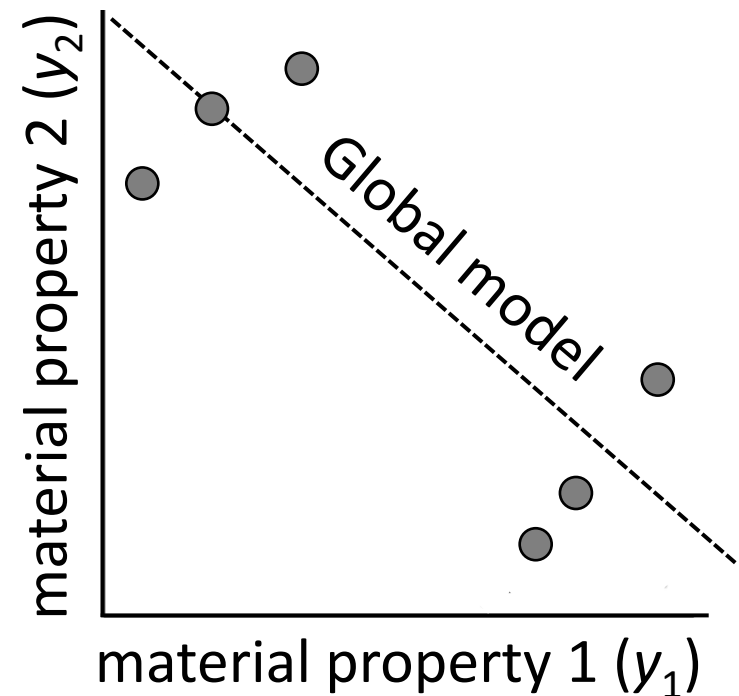


Feature selection
via **LASSO**+ l_0 [1]

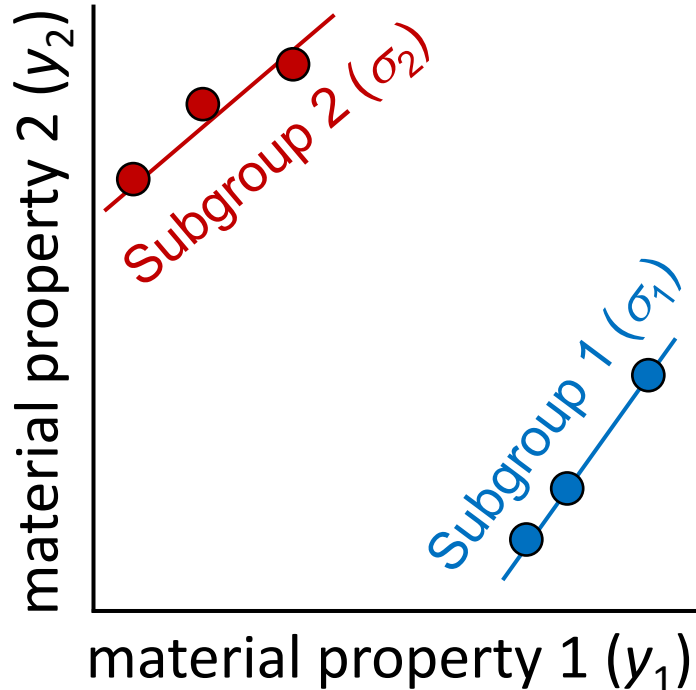
Typically one focuses on the inference of a global prediction model for some property of interest (e.g., LASSO, KRR, l_0)

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



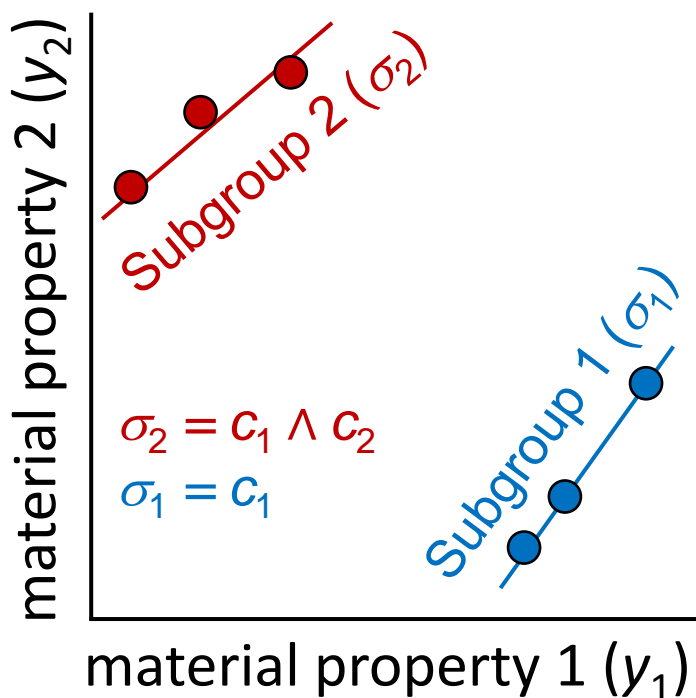
Periodic Table of the Elements																		18 VIIIA 8A			
1 1A H Hydrogen 1.008																	2 2A He Helium 4.003				
3 3A Li Lithium 6.941	4 4A Be Beryllium 9.012															5 5A B Boron 10.811	6 6A C Carbon 12.011	7 7A N Nitrogen 14.007	8 8A O Oxygen 15.999	9 9A F Fluorine 18.998	10 10A Ne Neon 20.180
11 1A Na Sodium 22.990	12 2A Mg Magnesium 24.305	13 3B Al Aluminum 26.982	14 4B Si Silicon 28.086	15 5B P Phosphorus 30.974	16 6B S Sulfur 32.065	17 7B Cl Chlorine 35.453	18 8B Ar Argon 39.948											19 9A K Potassium 39.098	20 10A Ca Calcium 40.078		
21 3B Sc Scandium 44.956	22 4B Ti Titanium 47.867	23 5B V Vanadium 50.942	24 6B Cr Chromium 51.996	25 7B Mn Manganese 54.938	26 8B Fe Iron 55.845	27 8B Co Cobalt 58.933	28 8B Ni Nickel 58.693	29 9B Cu Copper 63.546	30 10B Zn Zinc 65.38	31 11B Ga Gallium 69.723	32 12B Ge Germanium 72.631	33 13B As Arsenic 74.922	34 14B Se Selenium 78.96	35 15B Br Bromine 79.904	36 16B Kr Krypton 83.798						
37 1A Rb Rubidium 85.468	38 2A Sr Strontium 87.62	39 3B Y Yttrium 88.906	40 4B Zr Zirconium 91.224	41 5B Nb Niobium 92.906	42 6B Mo Molybdenum 95.94	43 7B Tc Technetium 98.907	44 8B Ru Ruthenium 101.07	45 8B Rh Rhodium 102.906	46 8B Pd Palladium 106.42	47 9B Ag Silver 107.868	48 10B Cd Cadmium 112.411	49 11B In Indium 114.818	50 12B Sn Tin 118.710	51 13B Sb Antimony 121.757	52 14B Te Tellurium 127.6	53 15B I Iodine 126.905	54 16B Xe Xenon 131.294				
55 1A Cs Cesium 132.905	56 2A Ba Barium 137.327	57-71 Lanthanide Series	72 5B Hf Hafnium 178.49	73 6B Ta Tantalum 180.948	74 7B W Tungsten 183.84	75 8B Re Rhenium 186.207	76 8B Os Osmium 190.23	77 8B Ir Iridium 192.225	78 9B Pt Platinum 195.084	79 10B Au Gold 196.967	80 11B Hg Mercury 200.592	81 12B Tl Thallium 204.383	82 13B Pb Lead 207.2	83 14B Bi Bismuth 208.980	84 15B Po Polonium [209]	85 16B At Astatine [210]	86 17B Rn Radon [222]				
87 1A Fr Francium [223]	88 2A Ra Radium [226]	89-103 Actinide Series	104 6B Rf Rutherfordium [261]	105 7B Db Dubnium [262]	106 8B Sg Seaborgium [266]	107 8B Bh Bohrium [264]	108 8B Hs Hassium [277]	109 9B Mt Meitnerium [268]	110 10B Ds Darmstadtium [271]	111 11B Rg Roentgenium [272]	112 12B Cn Copernicium [285]	113 13B Nh Nihonium [284]	114 14B Fl Flerovium [289]	115 15B Uup Ununpentium [288]	116 16B Lv Livermorium [293]	117 17B Uus Ununseptium [294]	118 18B Uuo Ununoctium [294]				
© 2015 Table International www.tableinternational.com																					

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, $a_1, \dots, a_m \in A$

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

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

e.g., HOMO-LUMO energy gap

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

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

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

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

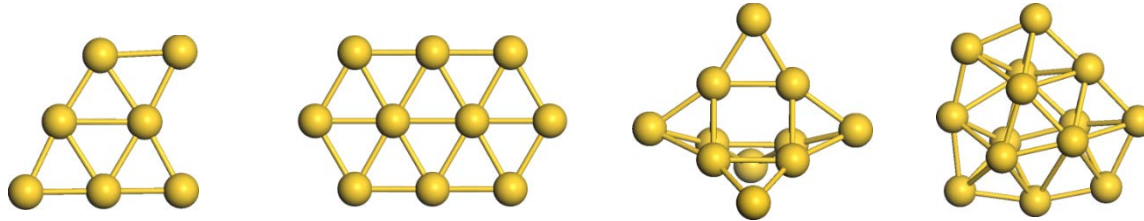
$\frac{|\text{ext}(\sigma)|}{|P|}$ is the coverage of points where σ is true

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

Two applications of subgroup discovery are presented here

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

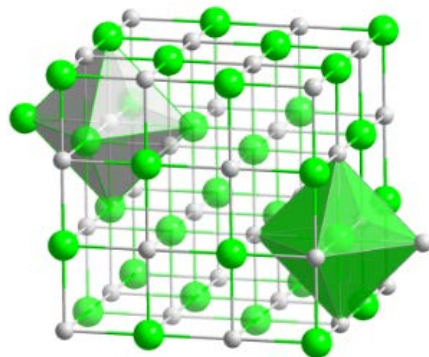
Display interesting optical, chemical, and electronic properties



24 400 gold cluster configurations in total

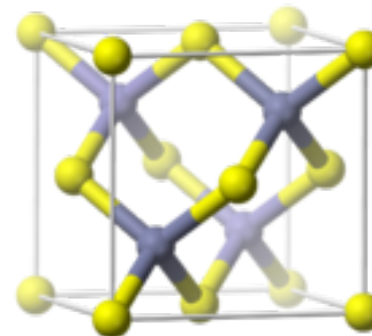
2. Classification of 82 Octet Binary Semiconductors

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

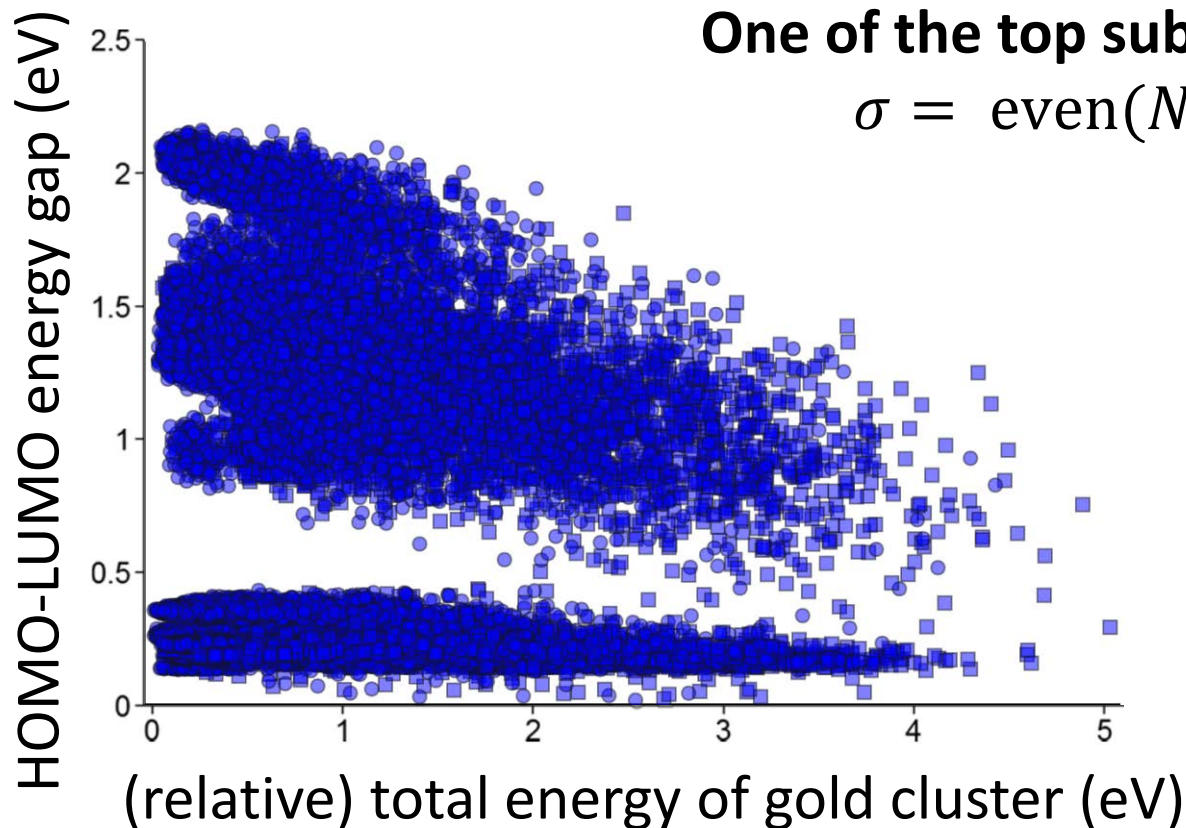
HOMO-LUMO energy gap

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

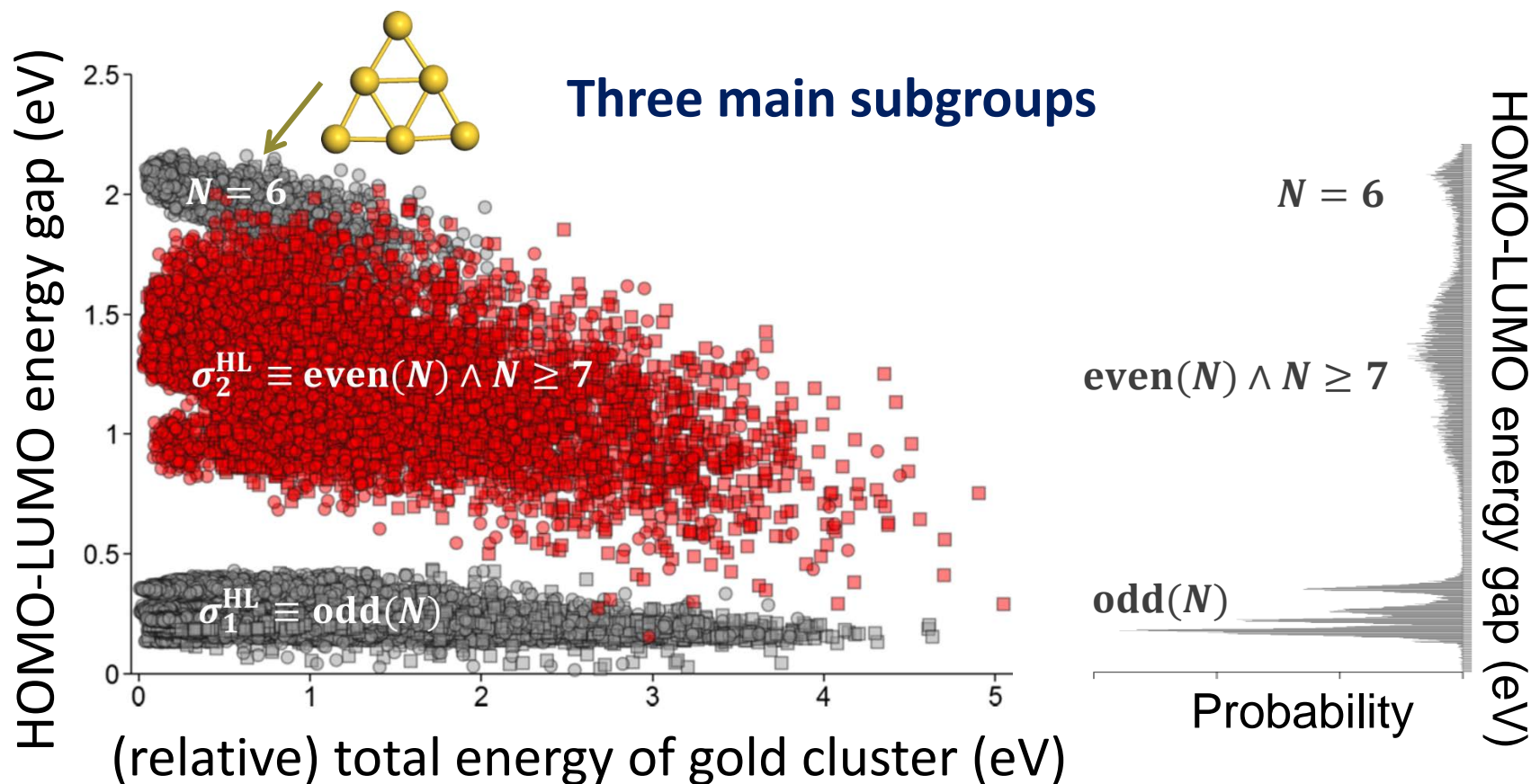
24 400 gold cluster configurations (of sizes 5-14) in the gas phase

Choose target property

HOMO-LUMO energy gap

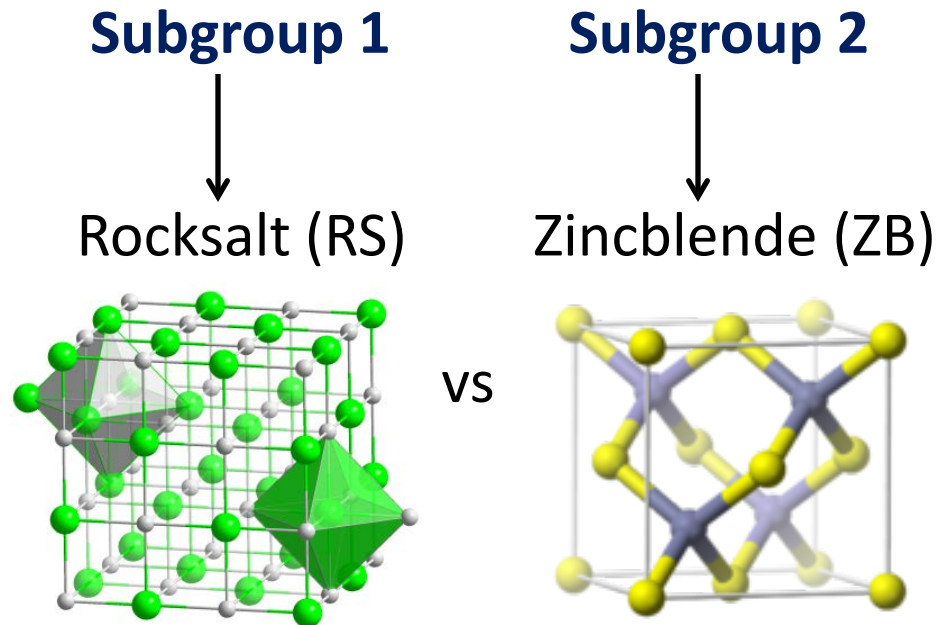
Choose variation reduction utility function

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

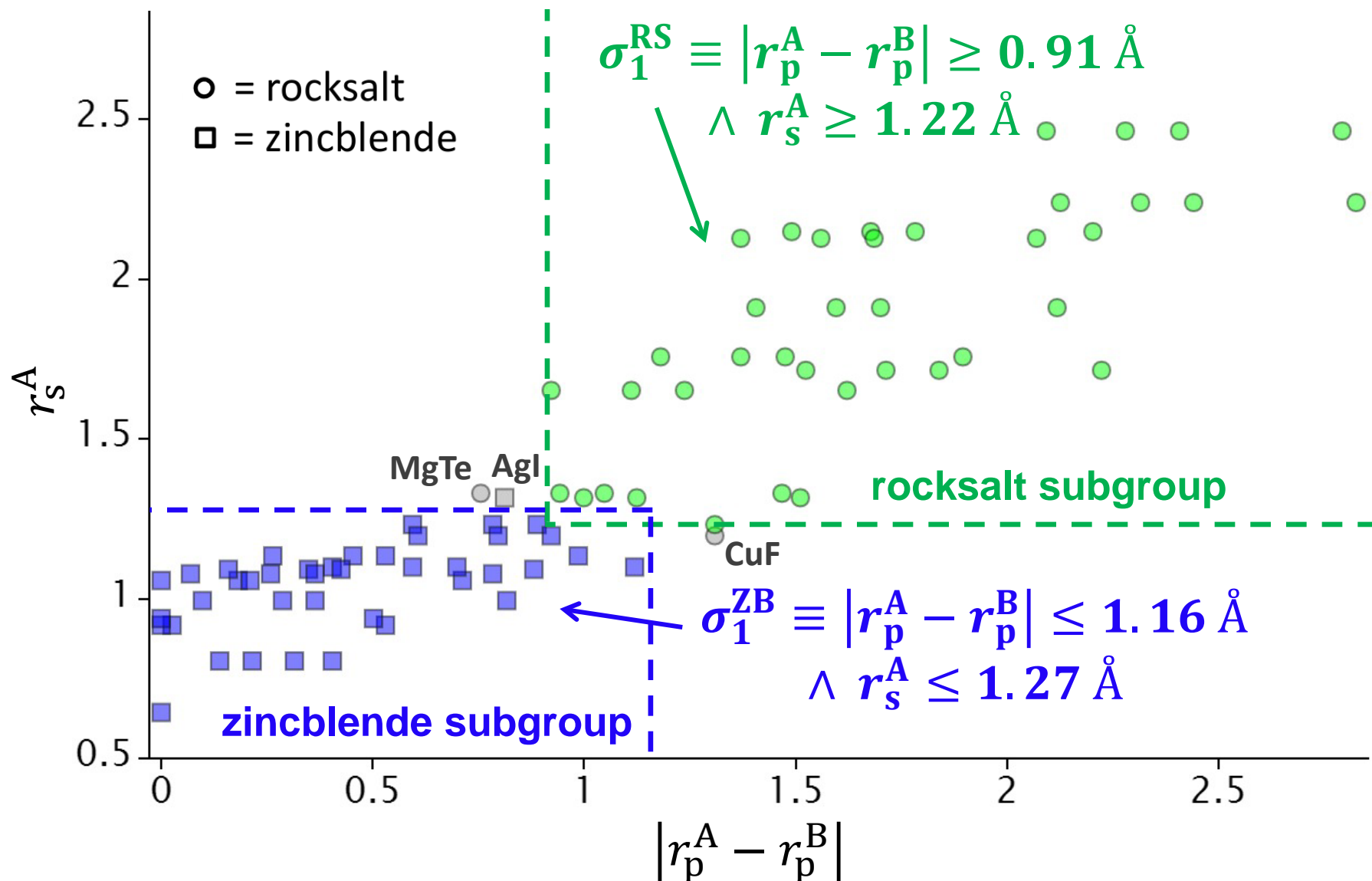


Can subgroup discovery find local models for the 82 octet binaries that describe zincblende and rocksalt?

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



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



Fritz Haber Institute of the Max Planck Society
Theory Department



Alexander von Humboldt
Stiftung / Foundation



mpi max planck institut
informatik