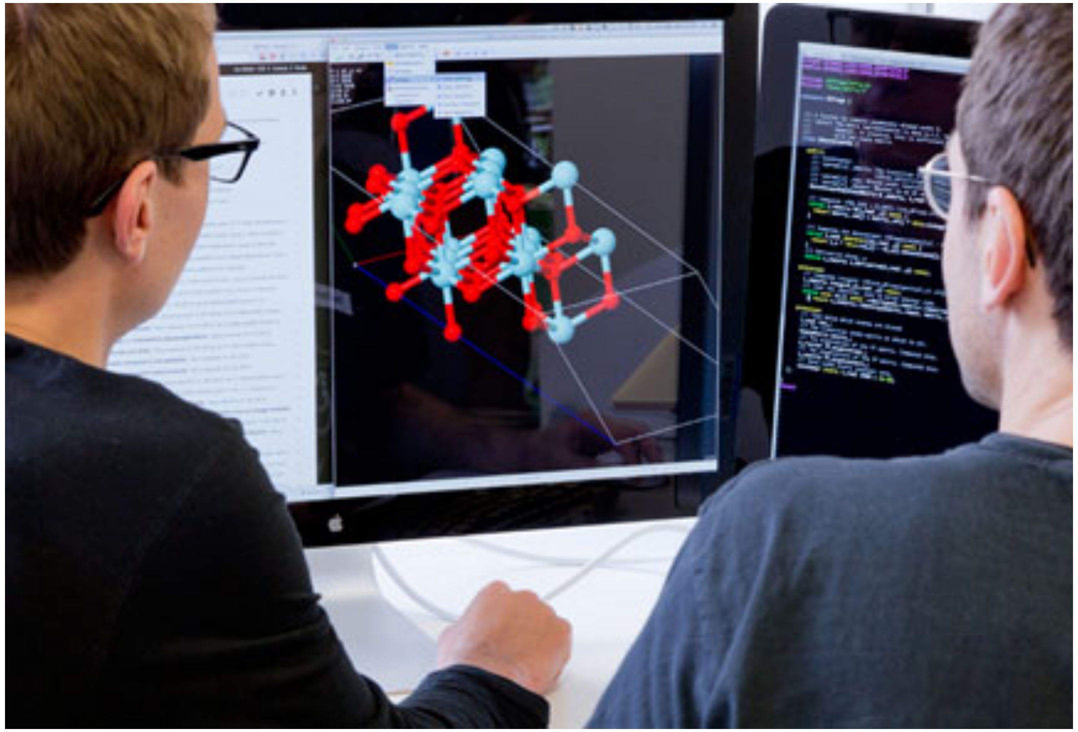
Introduction to Subgroup Discovery

NOMAD SUMMER
Lausanne, Sep 26 2018

Dr. Mario Boley
Max Planck Institute for Informatics
mboley@mpi-inf.mpg.de

Two flavors of data science

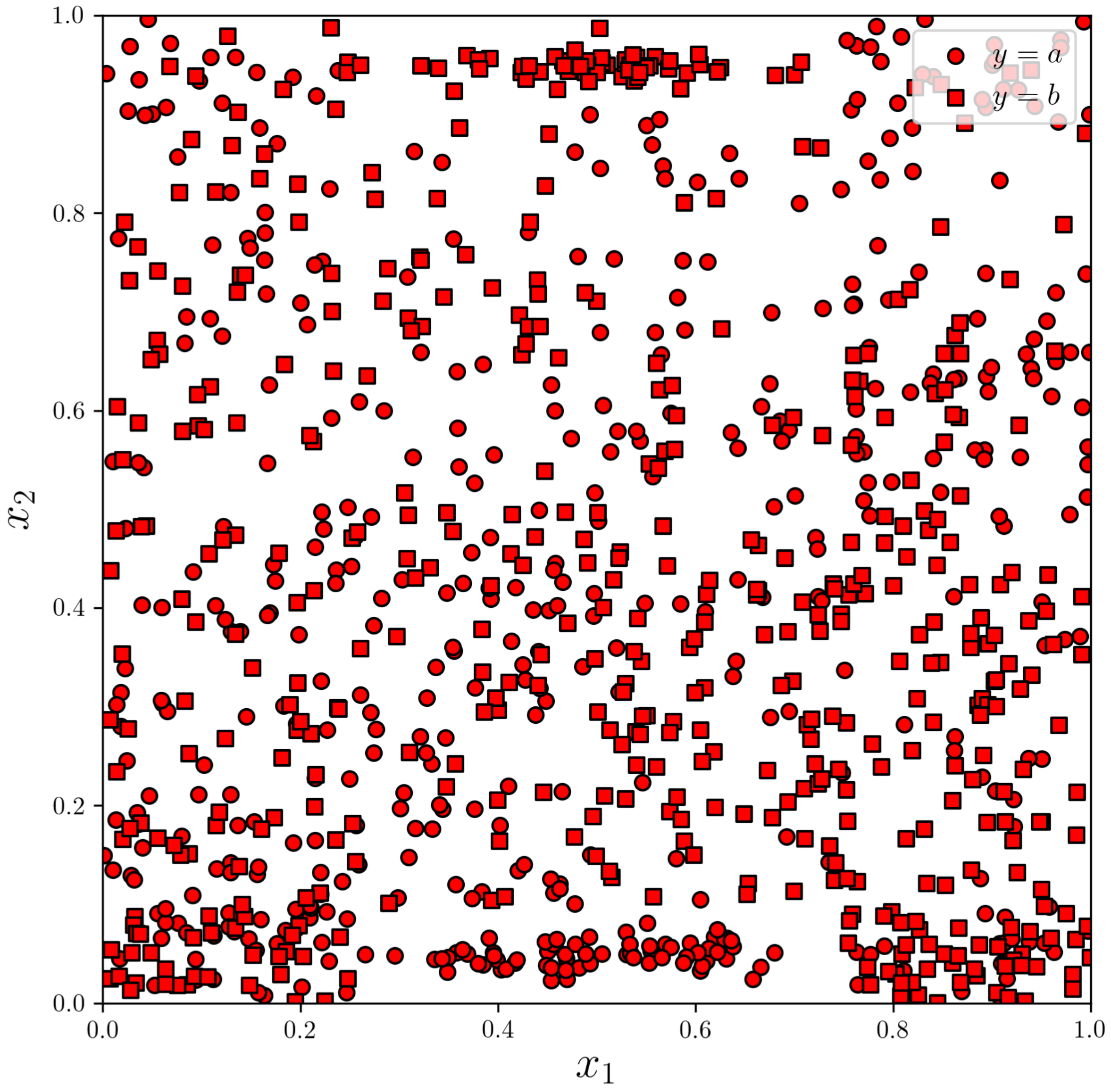
Predictive modelling



Data analytics

Basic setup

Given
Sample $S \subseteq P$
Target variable $y: P \rightarrow \{a, b, c, \dots\}$
Features $x_j: P \rightarrow X_j$



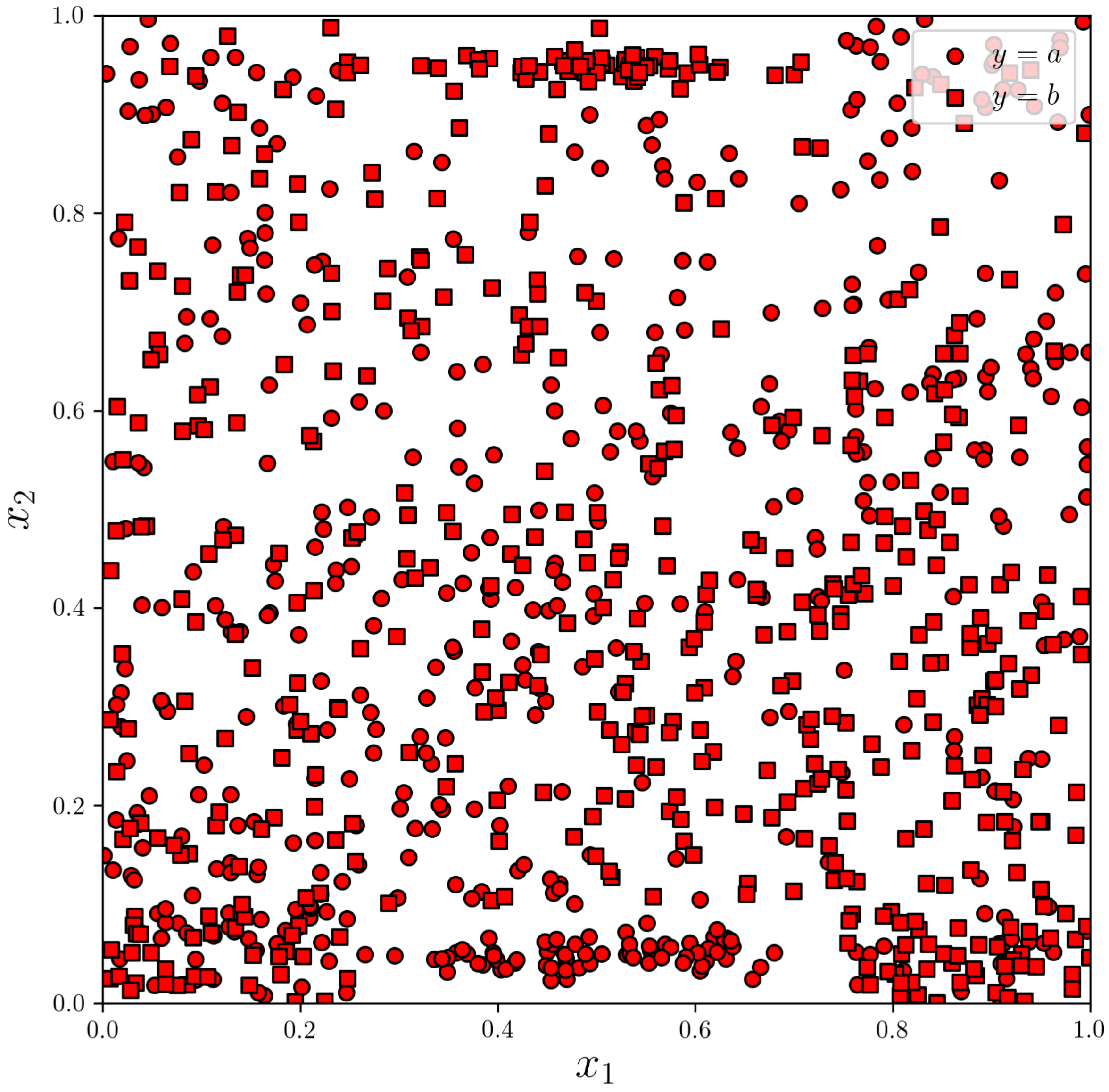
Basic setup

Given
Sample $S \subseteq P$
Target variable $y\colon P \to \{a, b, c, \dots\}$
Features $x_j\colon P \to X_j$

Predictive modeling
Model class $M \subseteq X \to Y$
Error measure $\mathrm{err}\colon Y \times Y \to \mathbb{R}_+$

Minimize

$$f(m) = \sum_{i \in S} \mathrm{err}\big(m(x(i)), y(i)\big)/|S| + \gamma \|m\|$$



Global models classify whole space

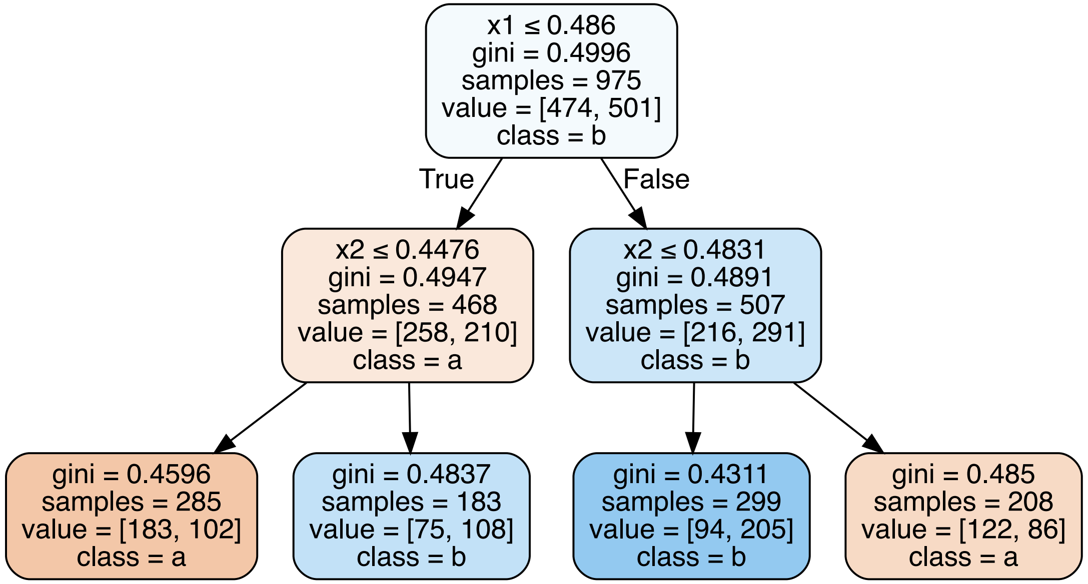
Given

Sample $S \subseteq P$

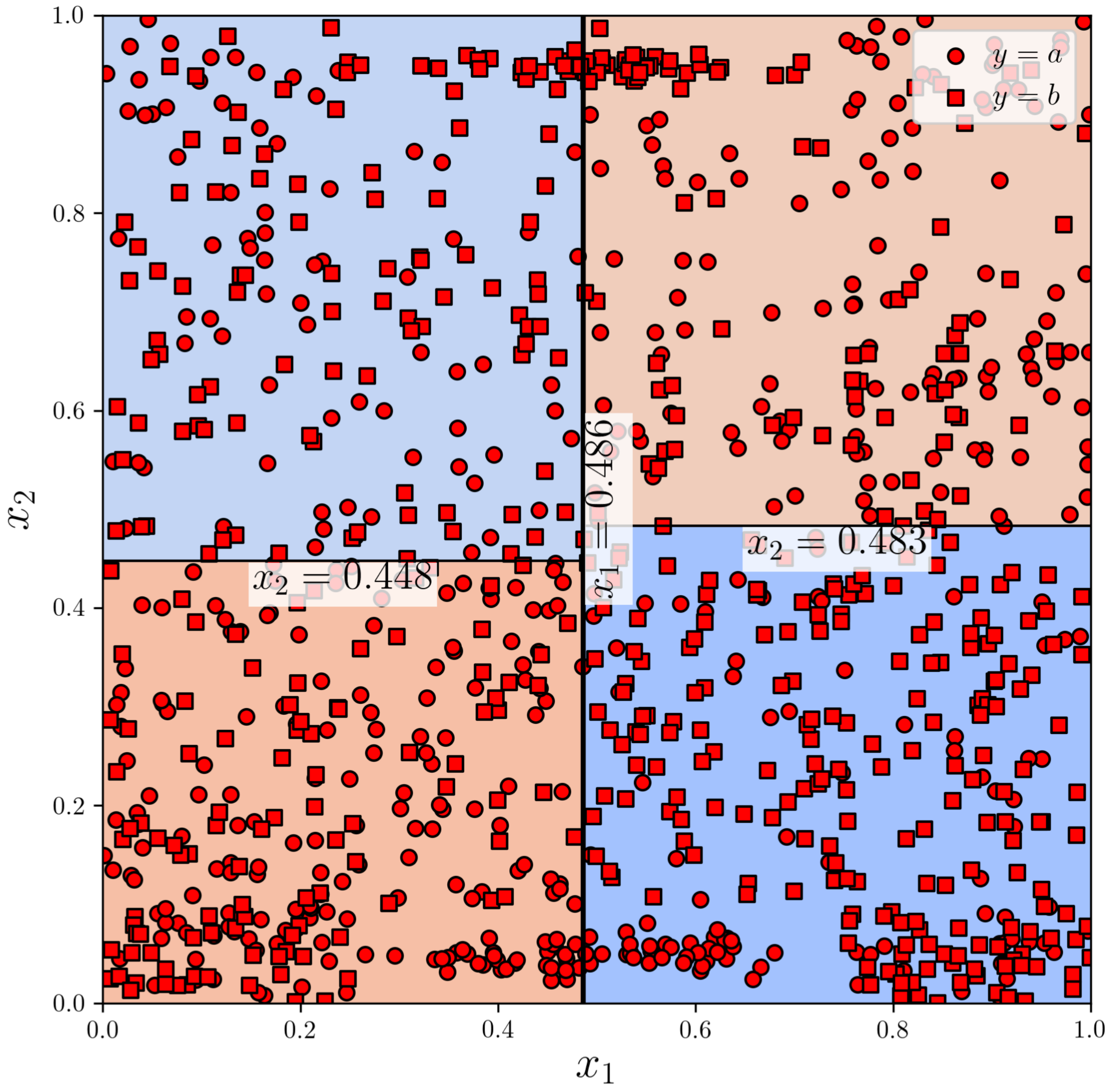
Target variable $y: P \to \{a, b, c, \ldots\}$

Features $x_j: P \to X_j$

Decision tree



Misses interesting local phenomena



Subgroup discovery focusses on local observations

Given

Sample $S \subseteq P$

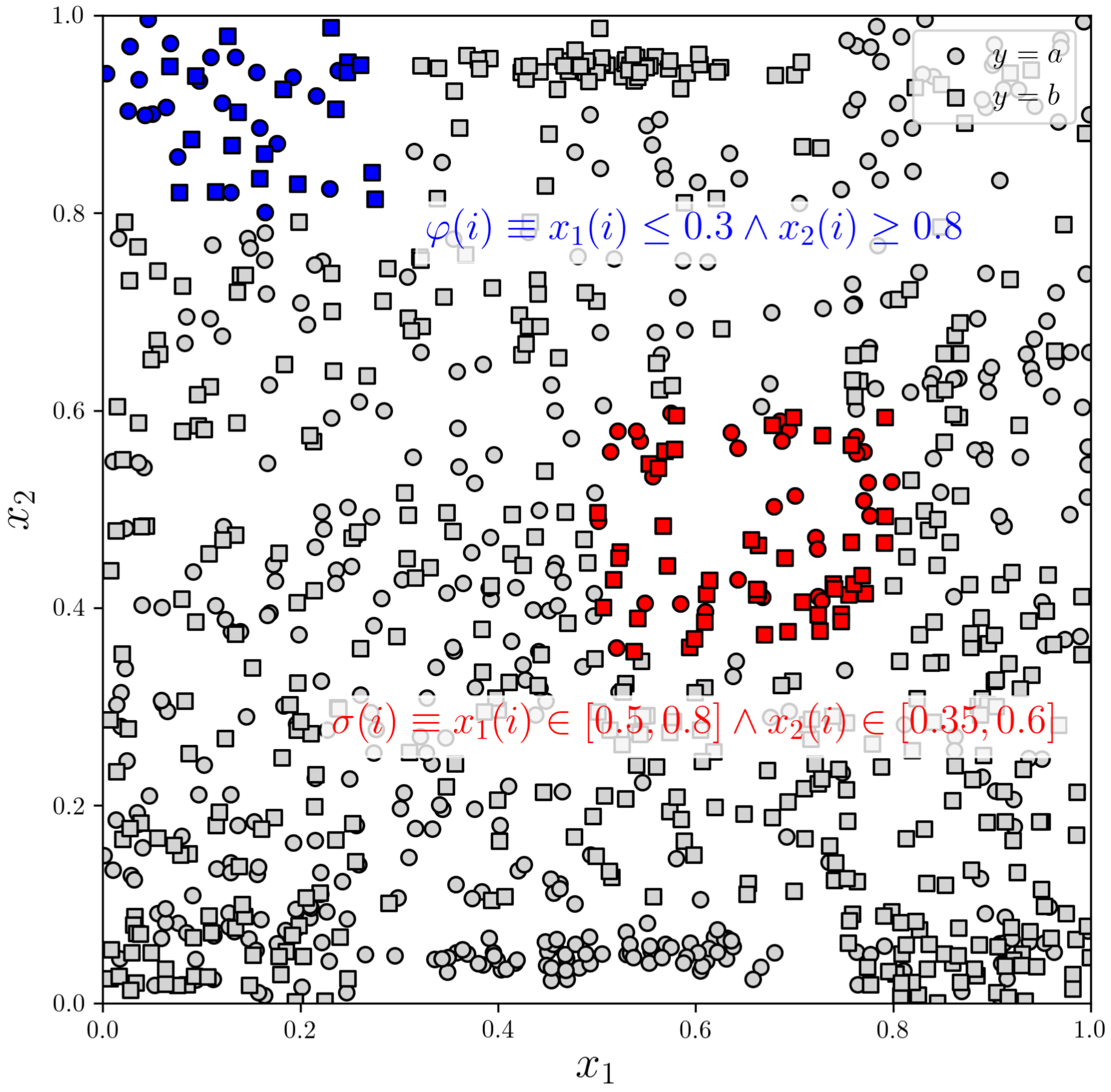
Target variable $y: P \rightarrow \{a, b, c, \ldots\}$

Features $x_j: P \rightarrow X_j$

Define

Propositions $\Pi_x = \{\pi_1, \ldots \pi_k\}$

Selection language $\mathcal{L}_x = \{\sigma(i) = \pi_{j_1}(i) \wedge \cdots \wedge \pi_{j_l}(i)\}$



Subgroup discovery focusses on local observations

Given

Sample $S \subseteq P$

Target variable $y: P \rightarrow \{a, b, c, \dots\}$

Features $x_j: P \rightarrow X_j$

Define

Propositions $\Pi_x = \{\pi_1, \dots \pi_k\}$

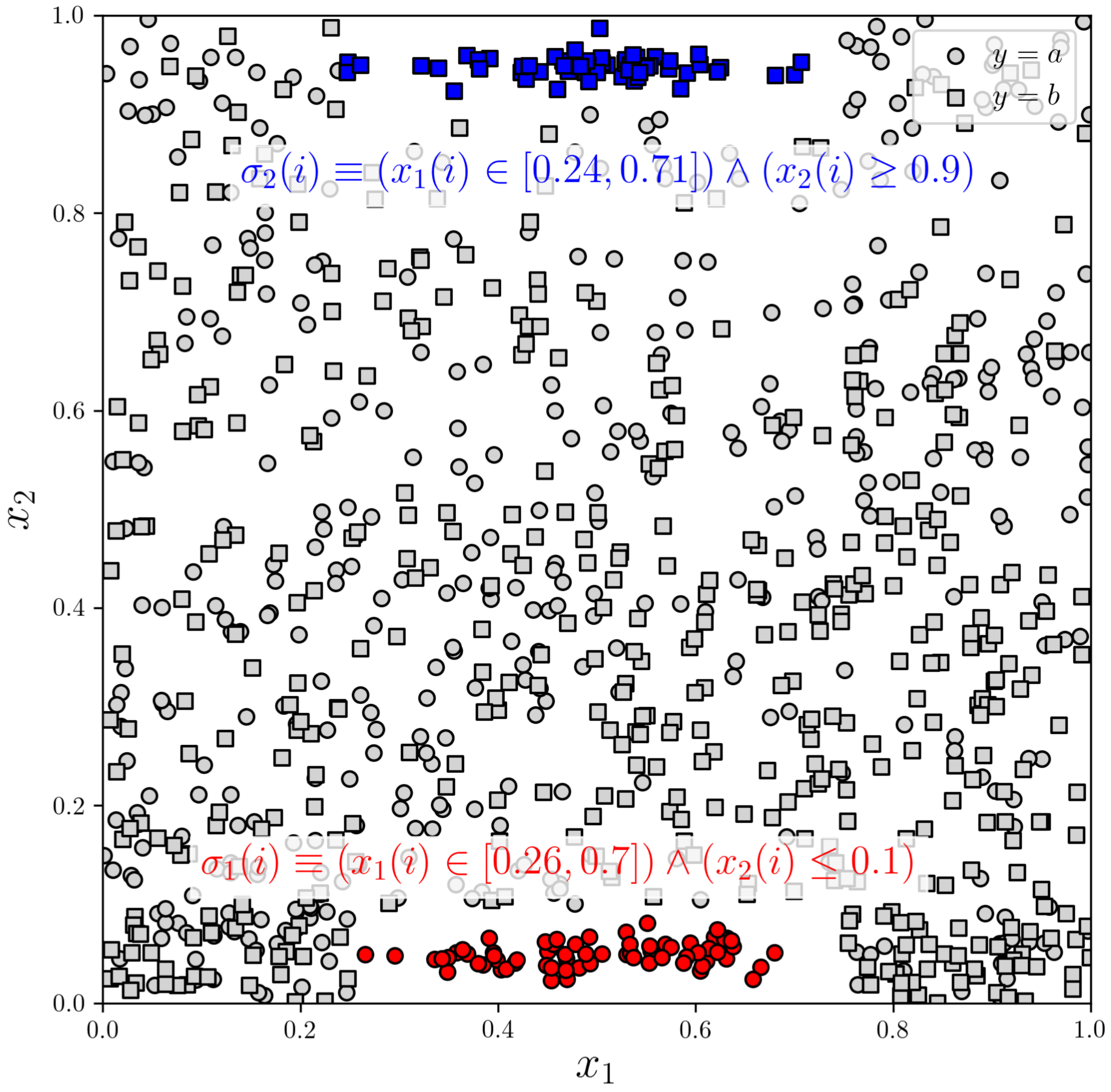
Selection language $\mathcal{L}_x = \left\{\sigma(i) = \pi_{j_1}(i) \wedge \cdots \wedge \pi_{j_l}(i)\right\}$

Optimize

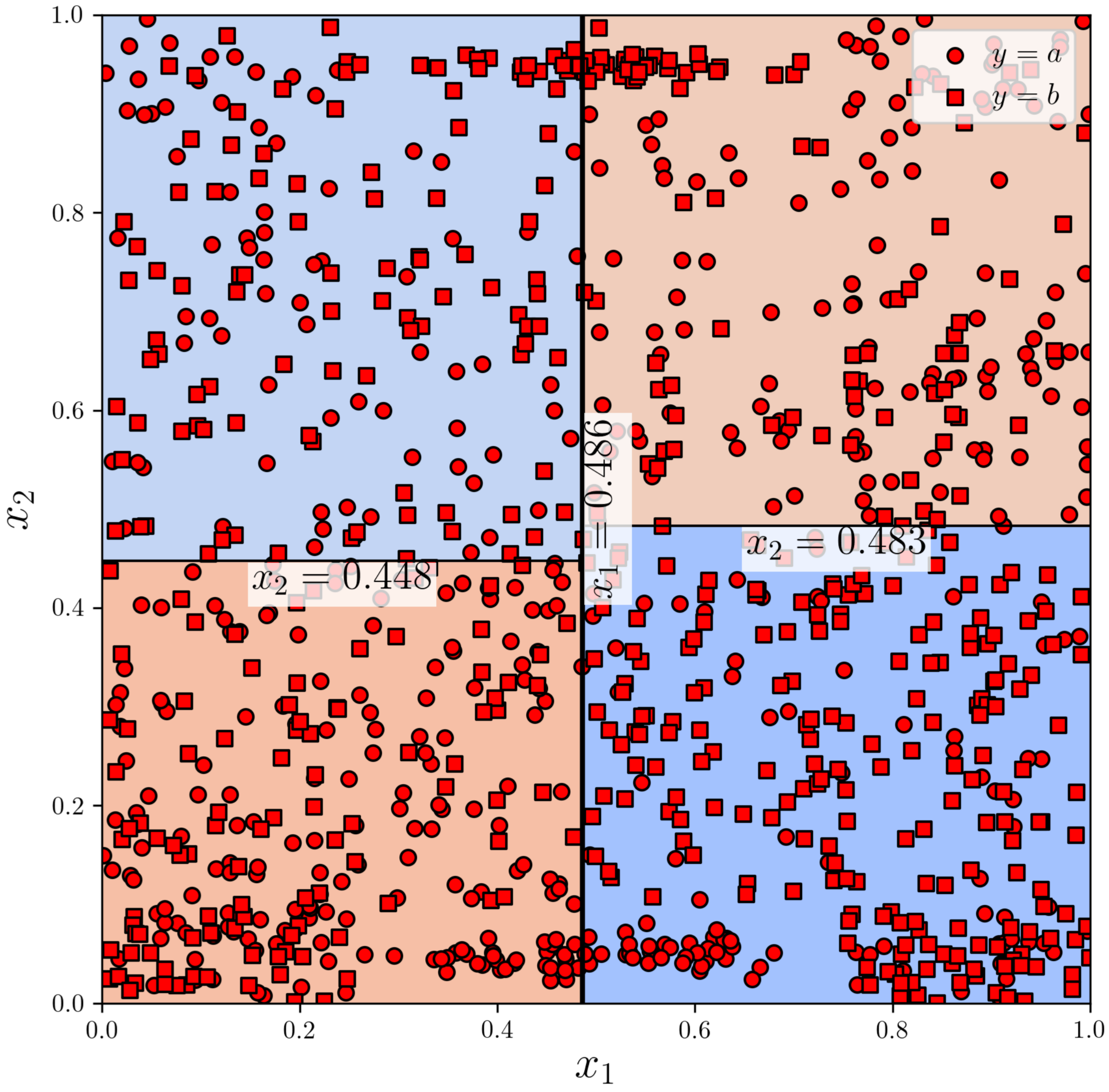
$$f(Q) = \mathrm{cov}(Q)^\gamma \mathrm{eff}(Q)_+$$

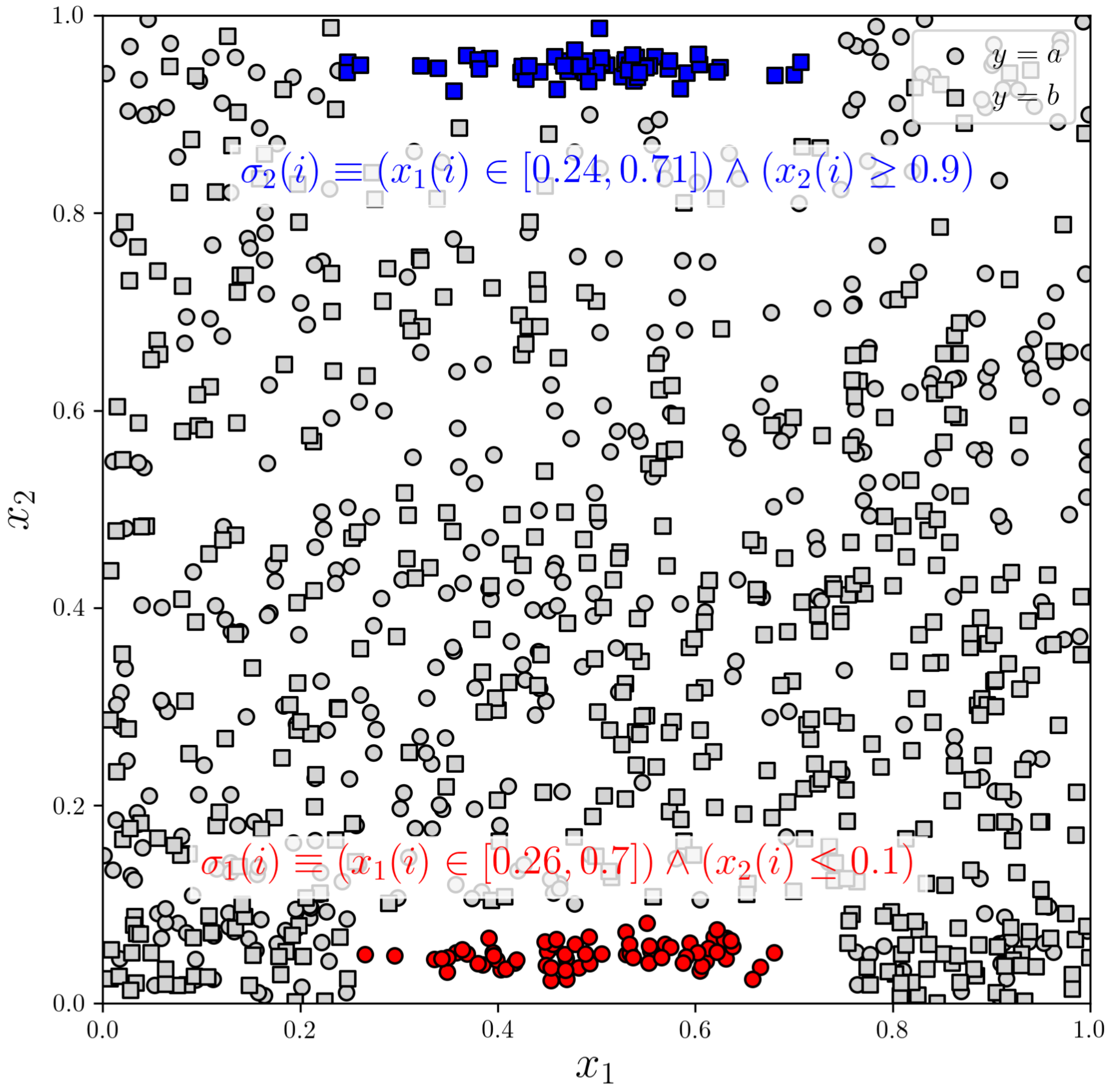
with

- $Q = \{i \in S: \sigma(i) = \top\}$ extension
- $\mathrm{cov}(Q) = |Q|/|S|$ coverage
- $\mathrm{eff}(Q) = \left(H_y(S) - H_y(Q)\right)/H_y(S)$ effect
- $H_y(Q) = -\sum_v p_Q(y = v) \log p_Q(y = v)$ entropy



Subgroup discovery focusses on local observations





Application 1: octet binary crystal structures

Population

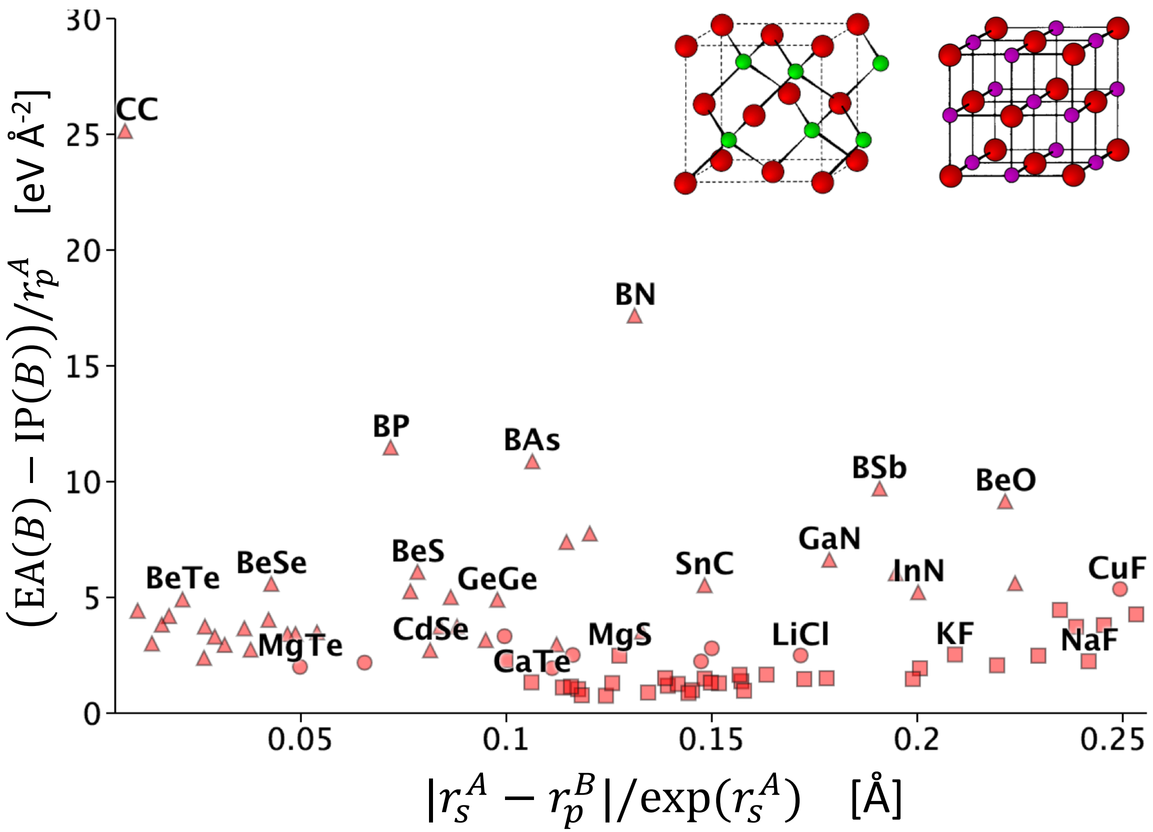
$P = \{AB : A = \mathrm{Ag}, \mathrm{Al}, \mathrm{Ba}, \ldots \wedge B = \mathrm{Br}, \mathrm{Cl}, \mathrm{F}, \ldots\}$

Target

$y = \mathrm{sign}(\Delta E)$ where $\Delta_E = E_{\mathrm{RS}} - E_{\mathrm{ZB}}$

Features

$x \in \{\mathrm{IP}^A, \mathrm{EA}^A, r_s^A, r_p^A, r_d^A, \mathrm{IP}^B, \mathrm{EA}^B, r_s^B, r_p^B, r_d^B,$
$\mathrm{IP}^A - \mathrm{IP}^B, \mathrm{EA}^A - \mathrm{EA}^B, |r_s^A - r_s^B|, \ldots\}$



[Ghiringhelli et al, PRL, 2015]

Application 1: octet binary crystal structures

Population

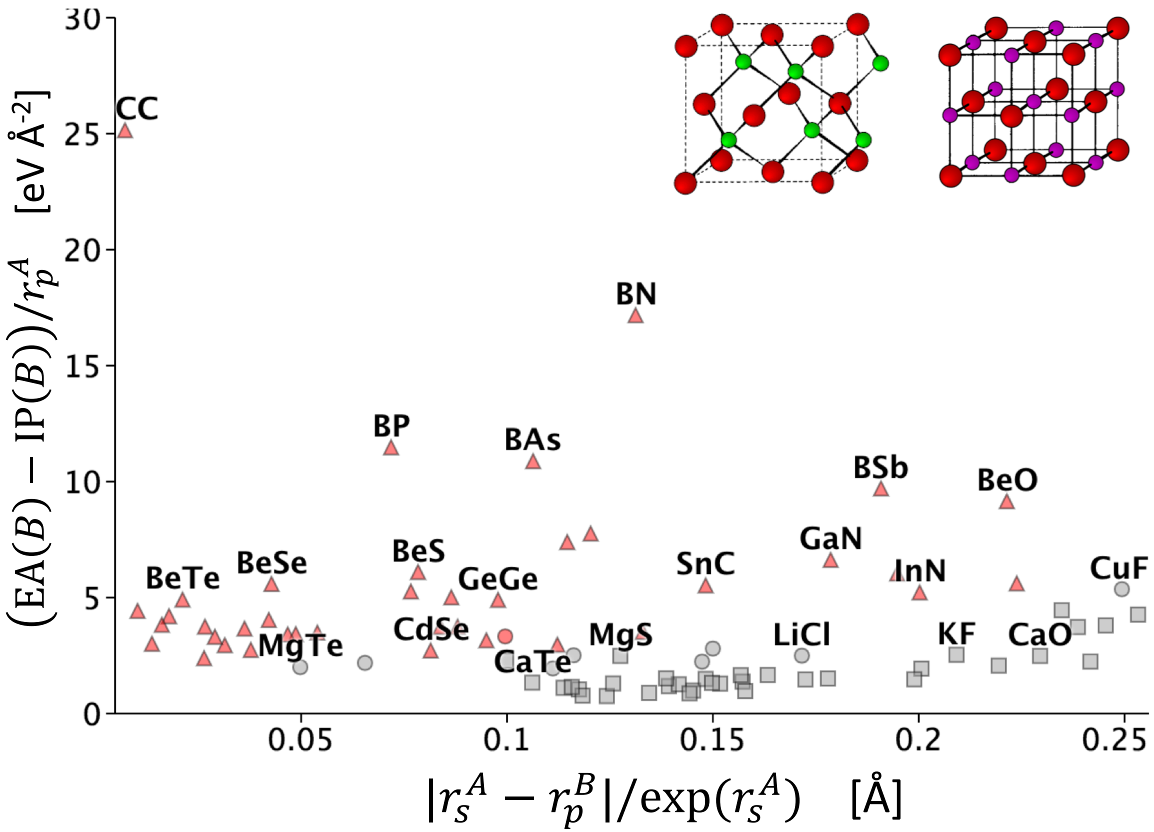
$P = \{AB: A = \text{Ag, Al, Ba, ...} \wedge B = \text{Br, Cl, F, ...}\}$

Target

$y = \text{sign}(\Delta E)$ where $\Delta_E = E_{\text{RS}} - E_{\text{ZB}}$

Features

$x \in \{\text{IP}^A, \text{EA}^A, r_s^A, r_p^A, r_d^A, \text{IP}^B, \text{EA}^B, r_s^B, r_p^B, r_d^B,$
$\text{IP}^A - \text{IP}^B, \text{EA}^A - \text{EA}^B, |r_s^A - r_s^B|, \dots\}$



Selector $\sigma_{\text{ZB}} \equiv (|r_p^A - r_p^B| \leq 1.15) \wedge (r_s^A \leq 1.27)$

Parameters $\text{cov} = 40/82$ $\text{eff} = 1$ $[H_y(\sigma_{\text{ZB}}) = 0, H_y(P) = 1]$

Application 1: octet binary crystal structures

Population

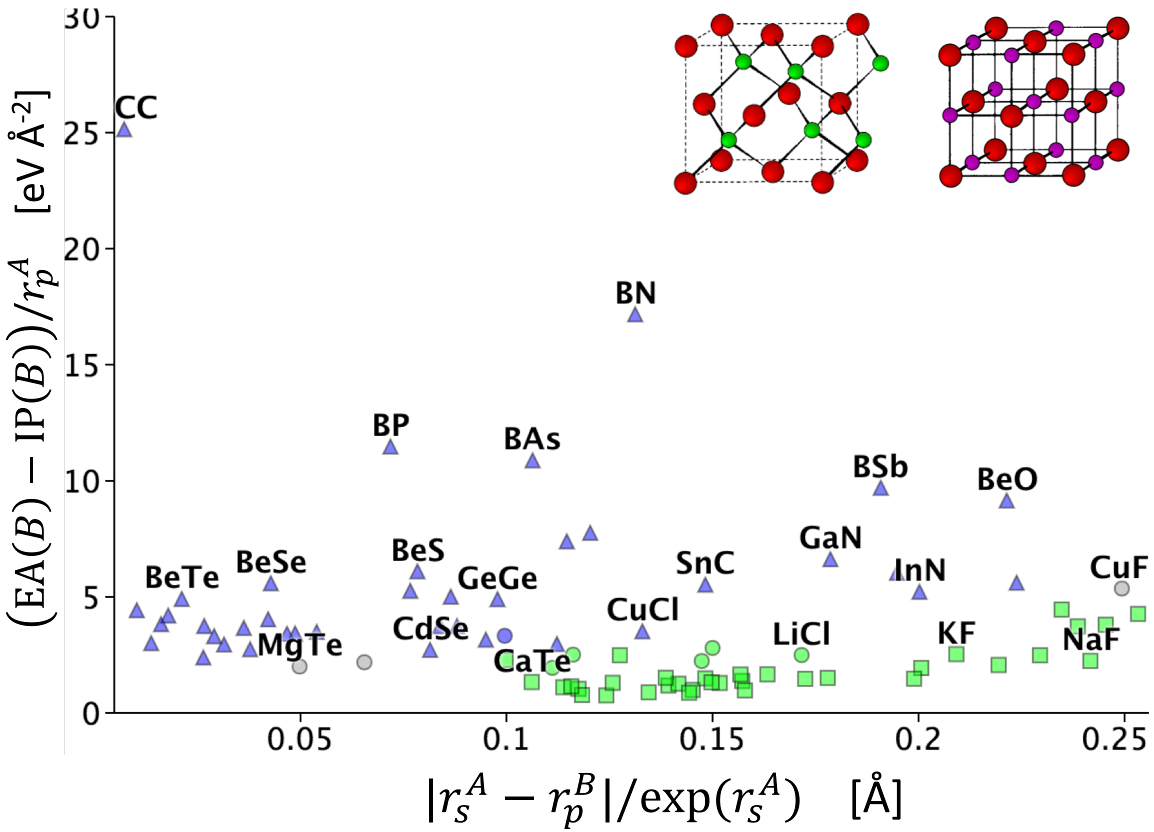
$$P = \{AB : A = \mathrm{Ag, Al, Ba}, \ldots \wedge B = \mathrm{Br, Cl, F}, \ldots\}$$

Target

$$y = \mathrm{sign}(\Delta E) \text{ where } \Delta_E = E_{\mathrm{RS}} - E_{\mathrm{ZB}}$$

Features

$$x \in \{\mathrm{IP}^A, \mathrm{EA}^A, r_s^A, r_p^A, r_d^A, \mathrm{IP}^B, \mathrm{EA}^B, r_s^B, r_p^B, r_d^B, \mathrm{IP}^A - \mathrm{IP}^B, \mathrm{EA}^A - \mathrm{EA}^B, |r_s^A - r_s^B|, \ldots\}$$



Selector $\sigma_{\mathrm{ZB}} \equiv (|r_p^A - r_p^B| \leq 1.15) \wedge (r_s^A \leq 1.27)$ $\sigma_{\mathrm{RS}} \equiv (|r_p^A - r_p^B| \geq 0.91) \wedge (r_s^A \geq 1.22)$

Parameters $\mathrm{cov} = 40/82$ $\mathrm{eff} = 1$ $[H_y(\sigma_{\mathrm{ZB}}) = 0, H_y(P) = 1]$

$\mathrm{cov} = 39/82$ $\mathrm{eff} = 1$ $[H_y(\sigma_{\mathrm{RS}}) = 0, H_y(P) = 1]$

Application 1: octet binary crystal structures

Population

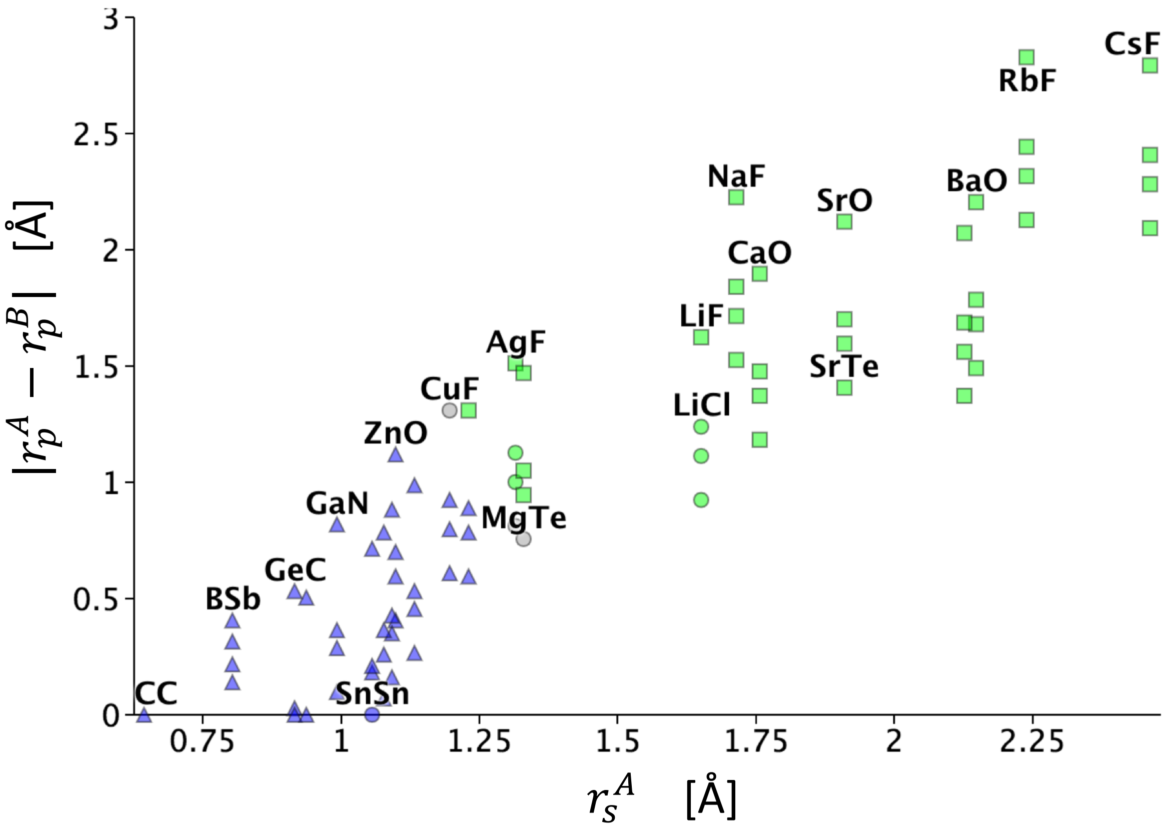
$$P = \{AB: A = \text{Ag}, \text{Al}, \text{Ba}, \ldots \wedge B = \text{Br}, \text{Cl}, \text{F}, \ldots\}$$

Target

$$y = \text{sign}(\Delta E) \text{ where } \Delta_E = E_{\text{RS}} - E_{\text{ZB}}$$

Features

$$x \in \{\text{IP}^A, \text{EA}^A, r_s^A, r_p^A, r_d^A, \text{IP}^B, \text{EA}^B, r_s^B, r_p^B, r_d^B, \text{IP}^A - \text{IP}^B, \text{EA}^A - \text{EA}^B, |r_s^A - r_s^B|, \ldots\}$$



Selector $\sigma_{\text{ZB}} \equiv \left(|r_p^A - r_p^B| \leq 1.15\right) \wedge (r_s^A \leq 1.27)$ $\sigma_{\text{RS}} \equiv \left(|r_p^A - r_p^B| \geq 0.91\right) \wedge (r_s^A \geq 1.22)$

Parameters $\text{cov} = 40/82$ $\text{eff} = 1$ $[H_y(\sigma_{\text{ZB}}) = 0, H_y(P) = 1]$

$\text{cov} = 39/82$ $\text{eff} = 1$ $[H_y(\sigma_{\text{RS}}) = 0, H_y(P) = 1]$

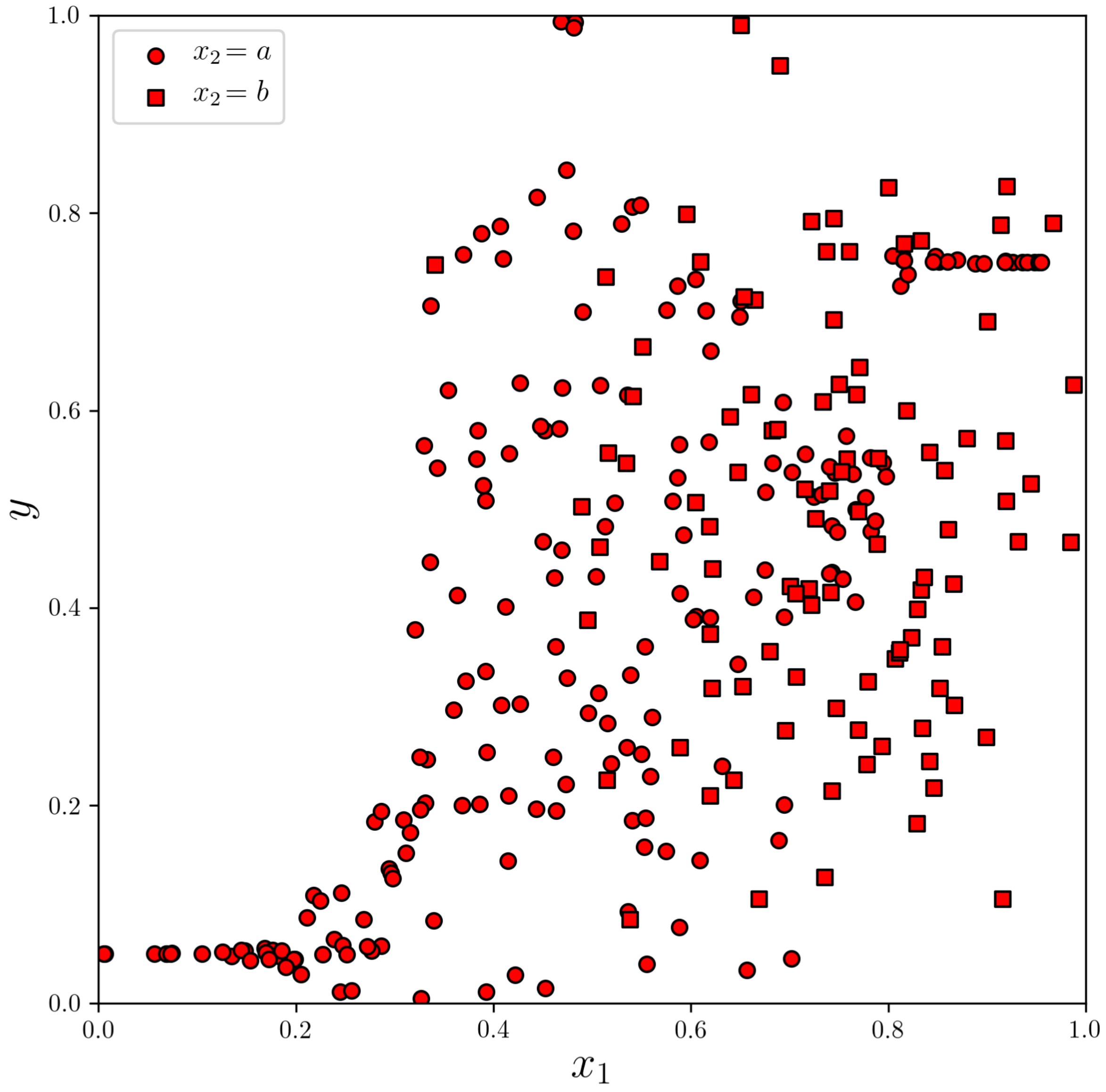
Subgroup discovery for real-valued targets

Given

Sample $S \subseteq P$

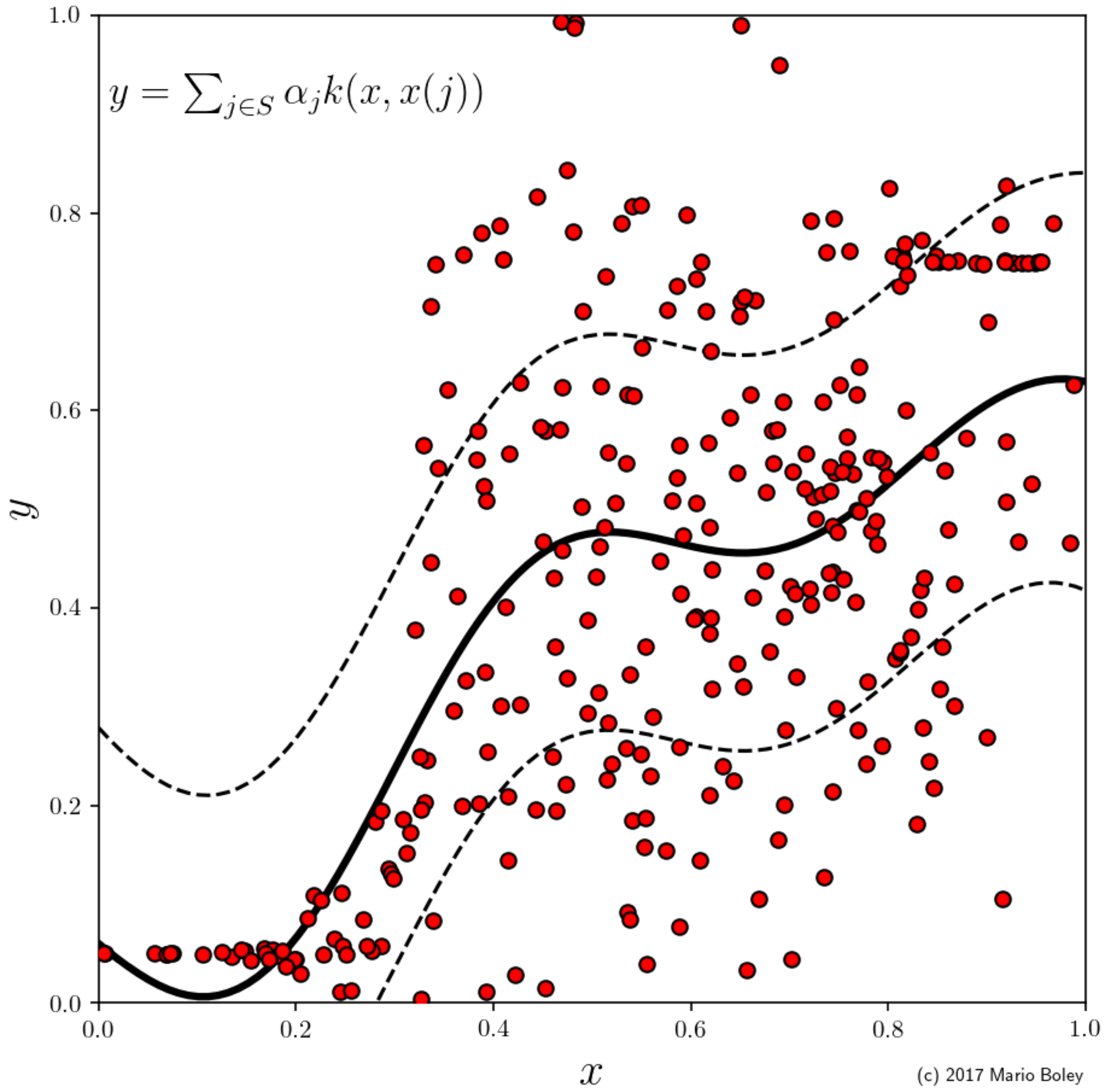
Target variable $y\colon P \to \mathbb{R}$

Features $x_j\colon P \to X_j$



Subgroup discovery for real-valued targets

Given
Sample $S \subseteq P$
Target variable $y: P \to \mathbb{R}$
Features $x_j: P \to X_j$



Subgroup discovery for real-valued targets

Given

Sample $S \subseteq P$

Target variable $y: P \rightarrow \mathbb{R}$

Features $x_j: P \rightarrow X_j$

Define

Propositions $\Pi_x = \{\pi_1, \dots \pi_k\}$

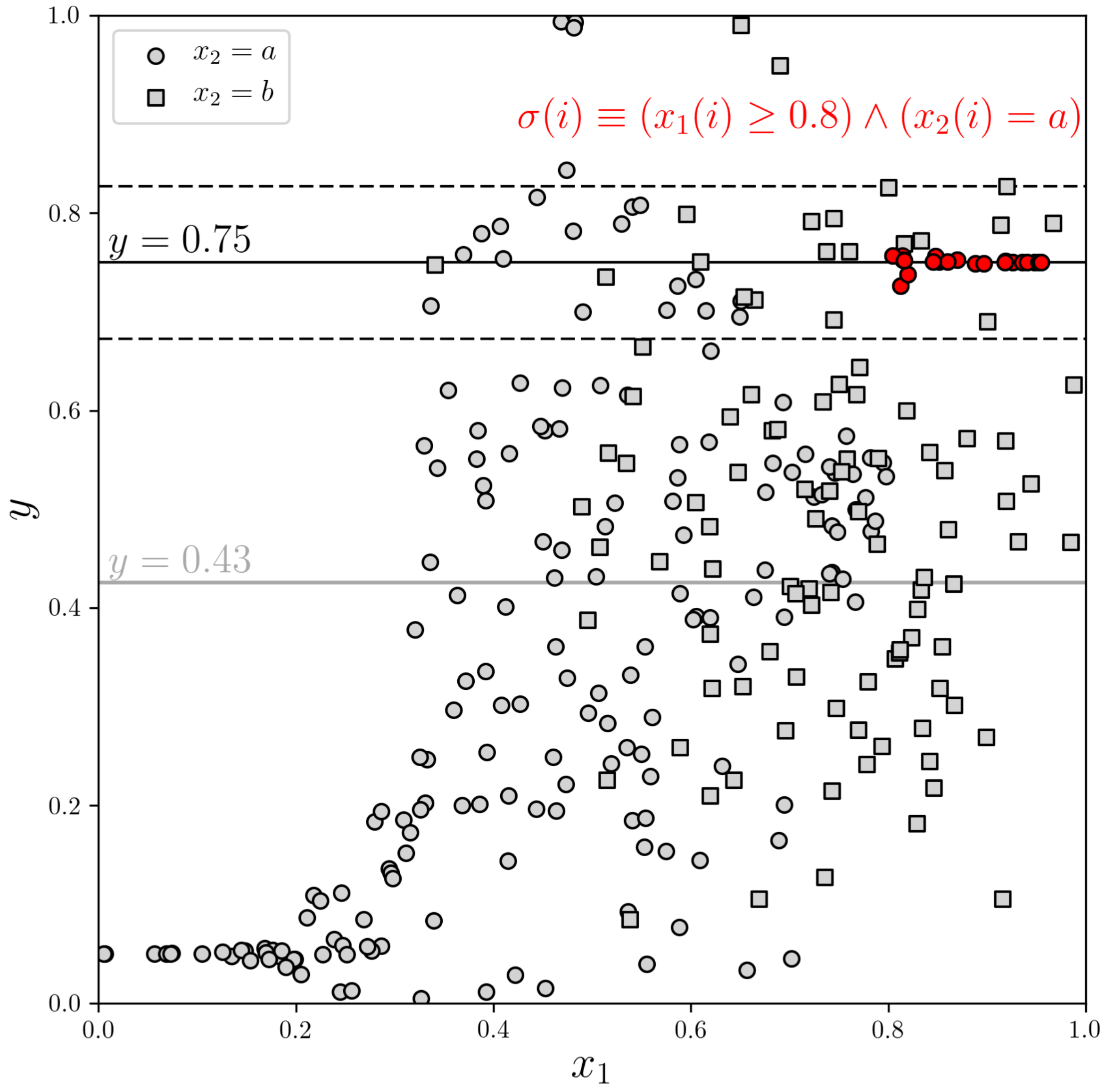
Selection language $\mathcal{L}_x = \{\sigma(i) = \pi_{j_1}(i) \wedge \dots \wedge \pi_{j_l}(i)\}$

Optimize

$$f(Q) = \mathrm{cov}(Q)^\gamma \mathrm{eff}(Q)_+$$

with

- $Q = \{i \in S: \sigma(i) = \top\}$ extension
- $\mathrm{cov}(Q) = |Q|/|S|$ coverage
- $\mathrm{eff}(Q) = \left(s_y(S) - s_y(Q)\right)/s_y(S)$ effect
- $s_y(Q) = \sqrt{\sum_{i \in Q} \left(\bar{y} - y(i)\right)^2 / (|Q| - 1)}$ std. dev.



Application 2: Au structure/property relationship

Population

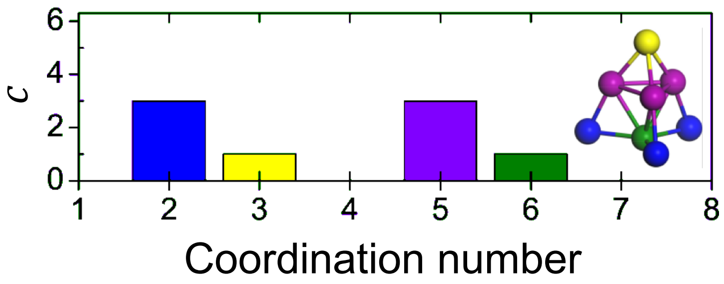
$$P = \{c : c \text{ conf. of Au5} - \text{Au14}\}$$

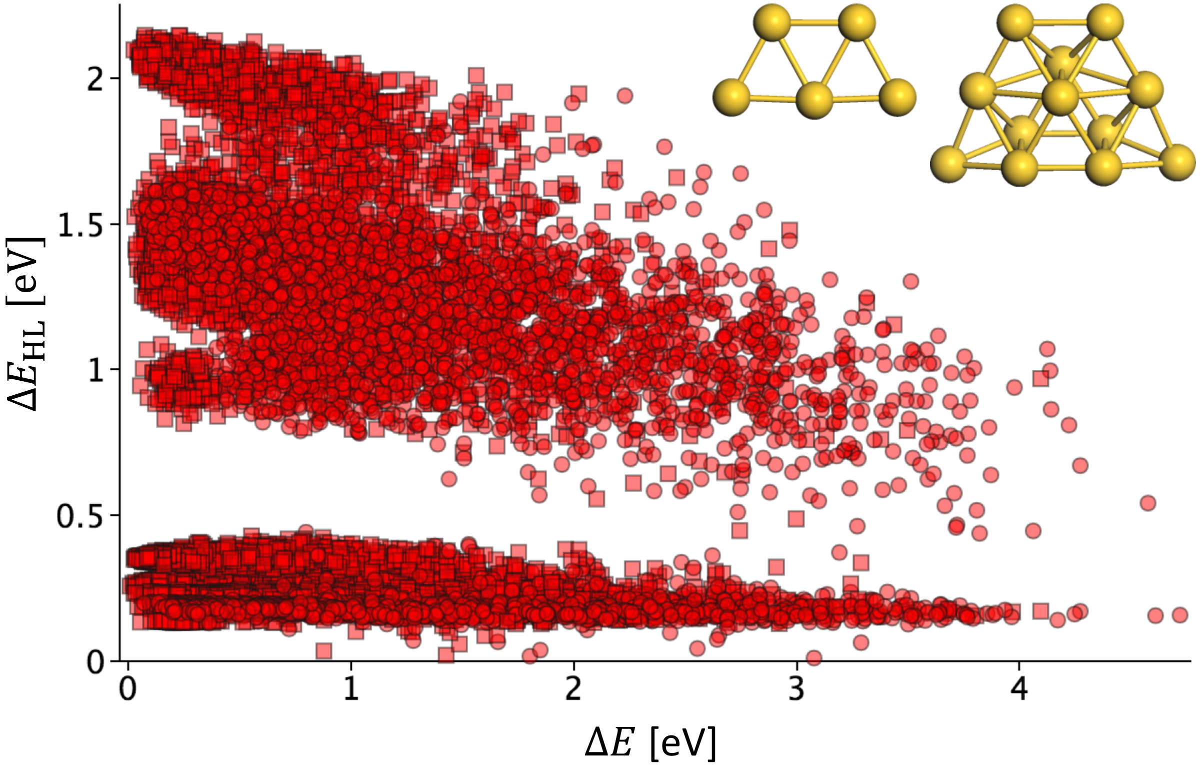
Target

$y = \Delta E_{\mathrm{HL}}$ HOMO-LUMO energy gap

Features

$$x \in \{a, c_1, c_2, c_3, c_4, c_5, c_6, r, \text{shape}, \mathrm{Mo_{co}}, \mathrm{Me_{co}}\}$$





Application 2: Au structure/property relationship

Population

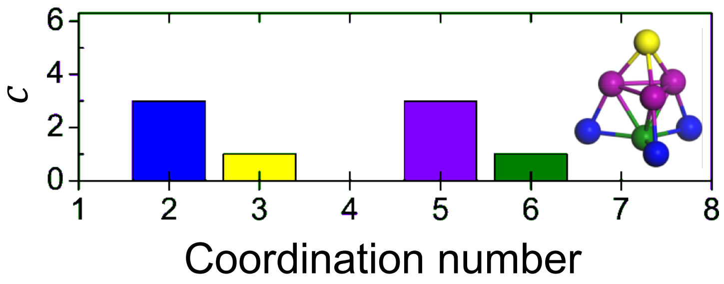
$$P = \{c : c \text{ conf. of Au5} - \text{Au14}\}$$

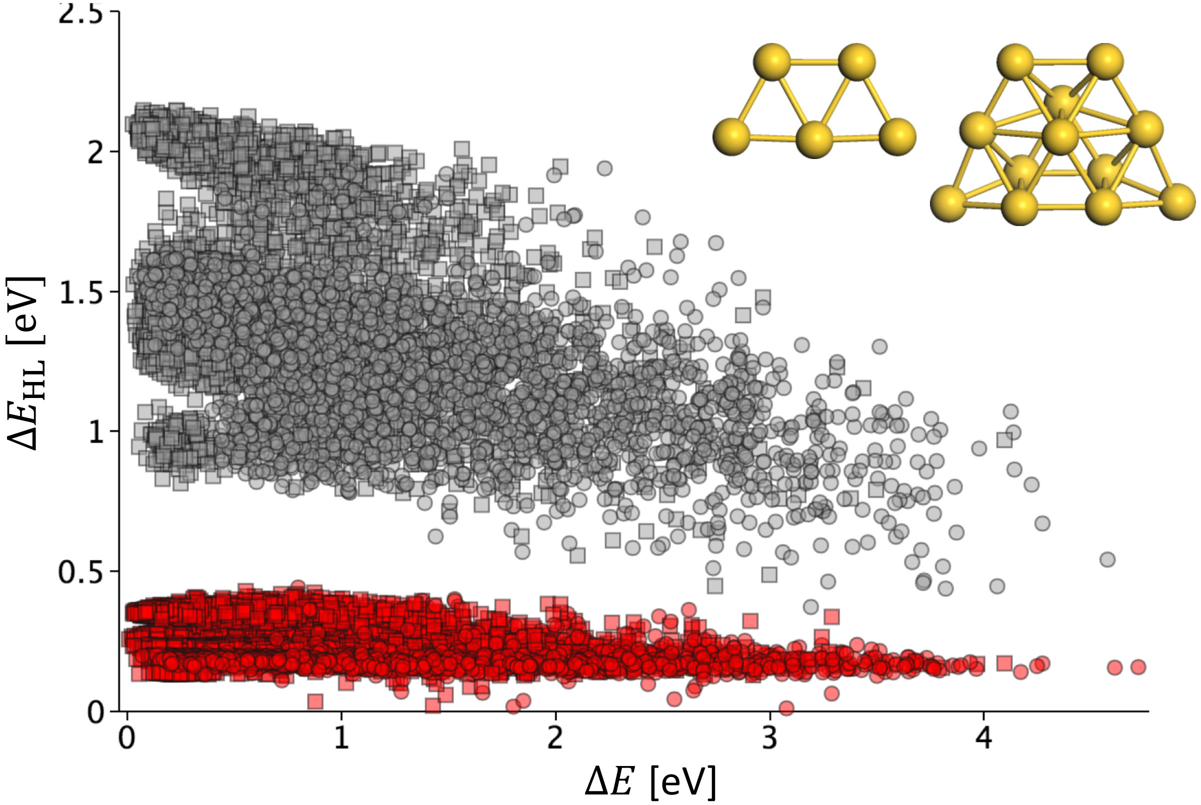
Target

$y = \Delta E_{\mathrm{HL}}$ HOMO-LUMO energy gap

Features

$$x \in \{a, c_1, c_2, c_3, c_4, c_5, c_6, r, \text{shape}, \mathrm{Mo}_{\mathrm{co}}, \mathrm{Me}_{\mathrm{co}}\}$$





Selector $\sigma(i) \equiv \mathrm{odd}(a(i))$

Parameters $\mathrm{cov}(\sigma) = 0.5$ $\mathrm{eff}(\sigma) = 0.9$ $[s_y(Q) = 0.06, s_y(S) = 0.58]$

$[\overline{y}(Q) = 0.22,\ \overline{y}(S) = 0.42]$

Application 2: Au structure/property relationship

Population

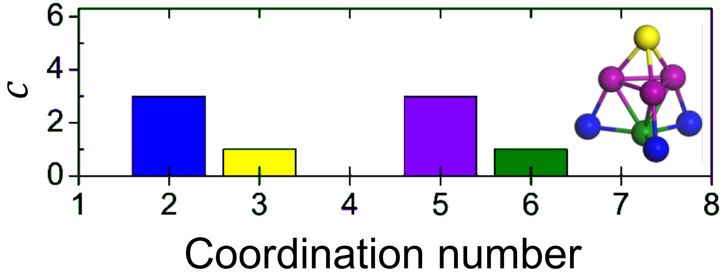
$$P = \{c : c \text{ conf. of Au5} - \text{Au14}\}$$

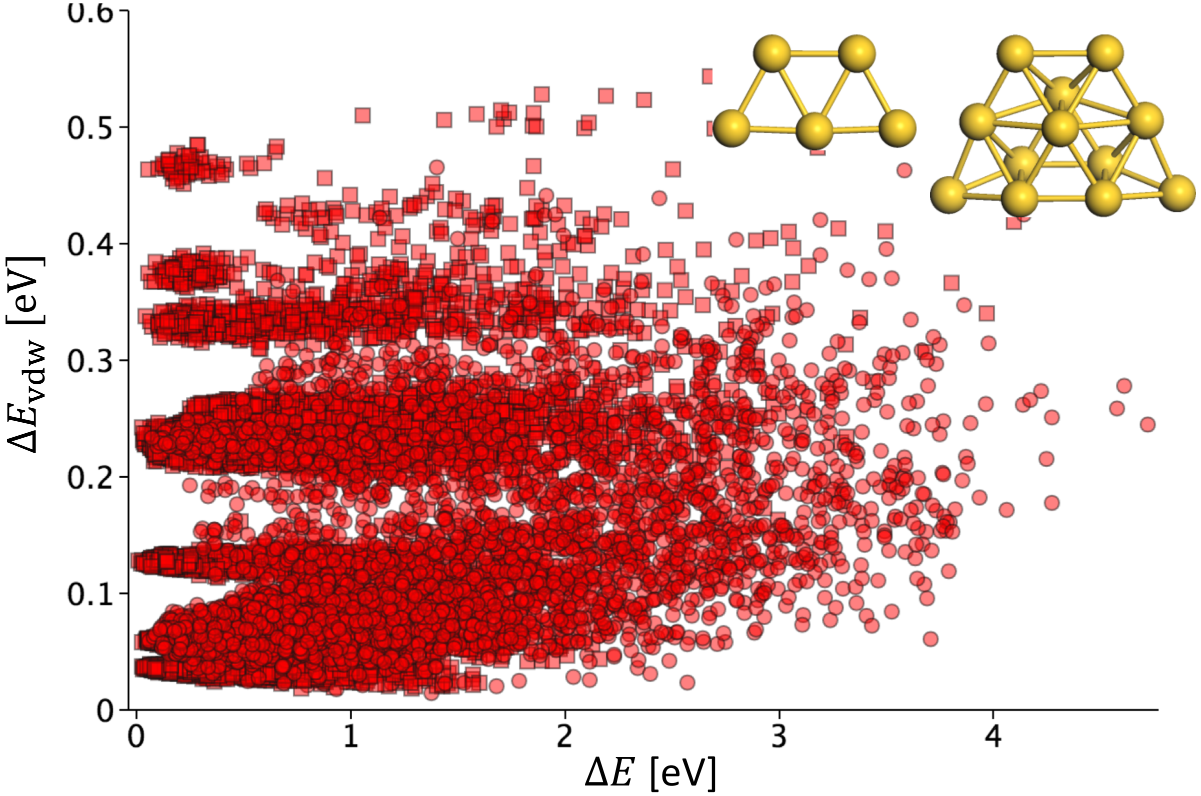
Target

$y = \Delta E_{\text{vdw}}$ van der Waals energy (ref.)

Features

$$x \in \{a, c_1, c_2, c_3, c_4, c_5, c_6, r, \text{shape}, \text{Mo}_{\text{co}}, \text{Me}_{\text{co}}\}$$





Application 2: Au structure/property relationship

Population

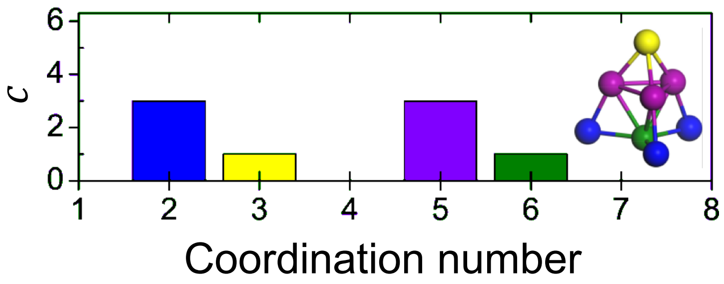
$$P = \{c: c \text{ conf. of Au5} - \text{Au14}\}$$

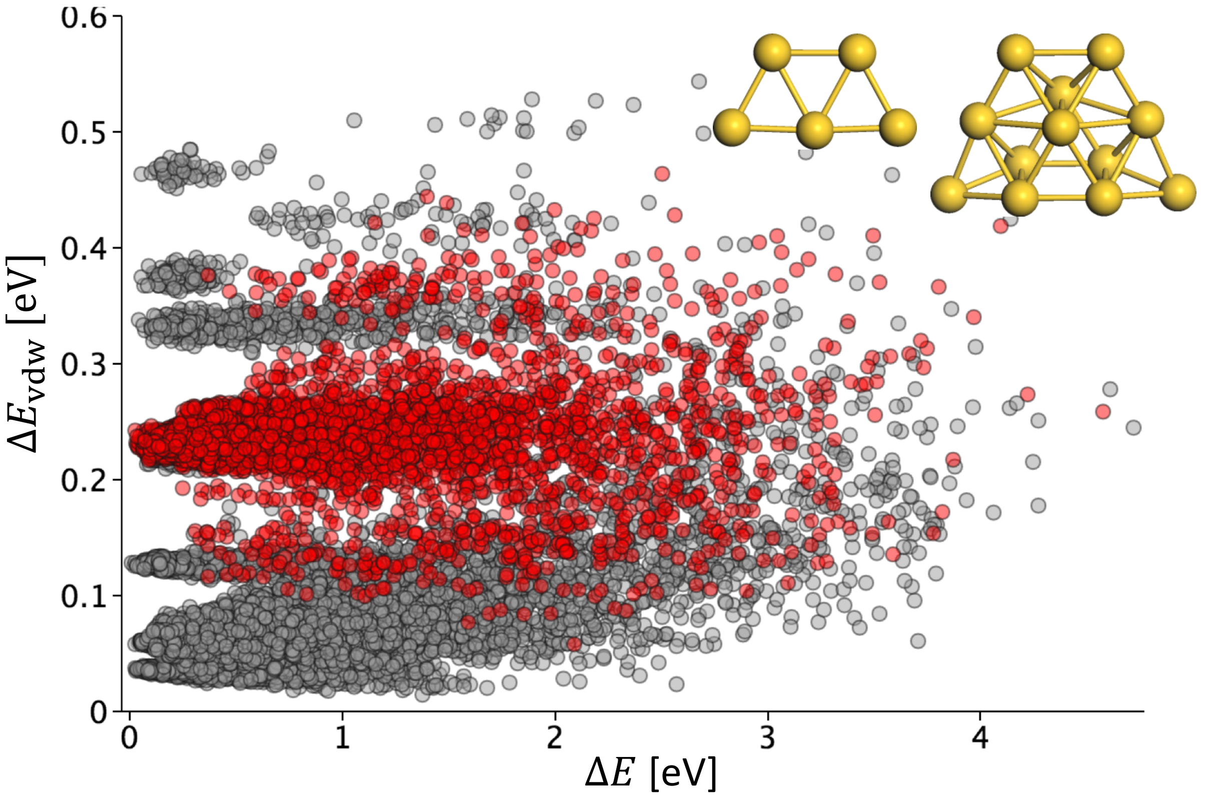
Target

$y = \Delta E_{\text{vdw}}$ van der Waals energy (ref.)

Features

$$x \in \{a, c_1, c_2, c_3, c_4, c_5, c_6, r, \text{shape}, \text{Mo}_{\text{co}}, \text{Me}_{\text{co}}\}$$





Selector $\sigma(i) \equiv a(i) \in [8,12] \wedge c_2(i) > 0.17 \wedge c_6(i) < 0.28 \wedge r(i) > 0.86$

Parameters $\text{cov}(\sigma) = 0.2$ $\text{eff}(\sigma) = 0.68$ $[s_y(Q) = 0.03,\ s_y(S) = 0.09]$

$[\bar{y}(Q) = 0.23,\ \bar{y}(S) = 0.13]$

Subgroup discovery with multiple targets

Given

Sample $S \subseteq P$

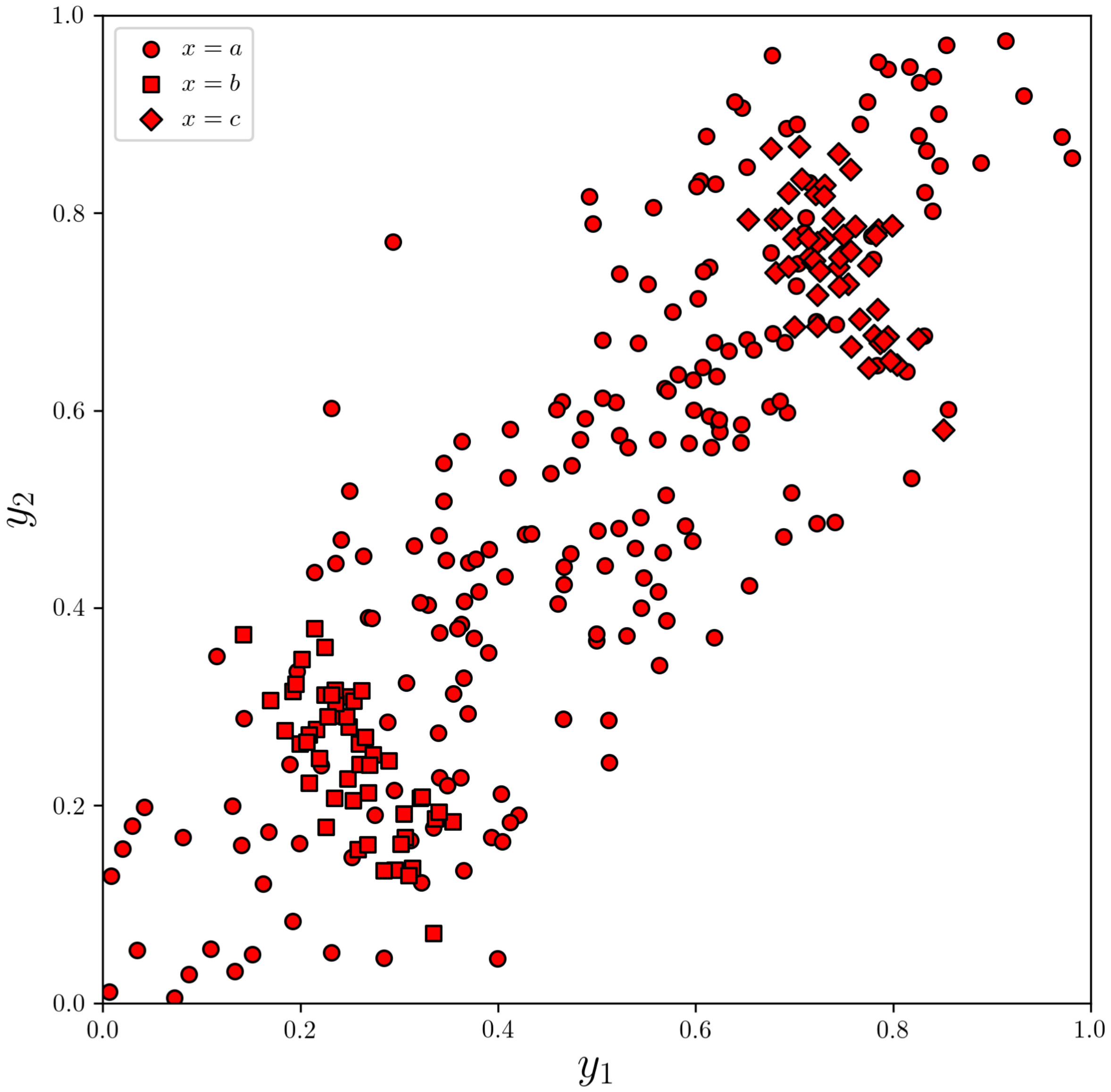
Target variable $y_1, y_2 \colon P \to \mathbb{R}$

Features $x_j \colon P \to X_j$

Define

Propositions $\Pi_x = \{\pi_1, \dots \pi_k\}$

Selection language $\mathcal{L}_x = \{\sigma(i) = \pi_{j_1}(i) \wedge \dots \wedge \pi_{j_l}(i)\}$



Subgroup discovery with multiple targets

Given

Sample $S \subseteq P$

Target variable $y_1, y_2 \colon P \to \mathbb{R}$

Features $x_j \colon P \to X_j$

Define

Propositions $\Pi_x = \{\pi_1, \dots \pi_k\}$

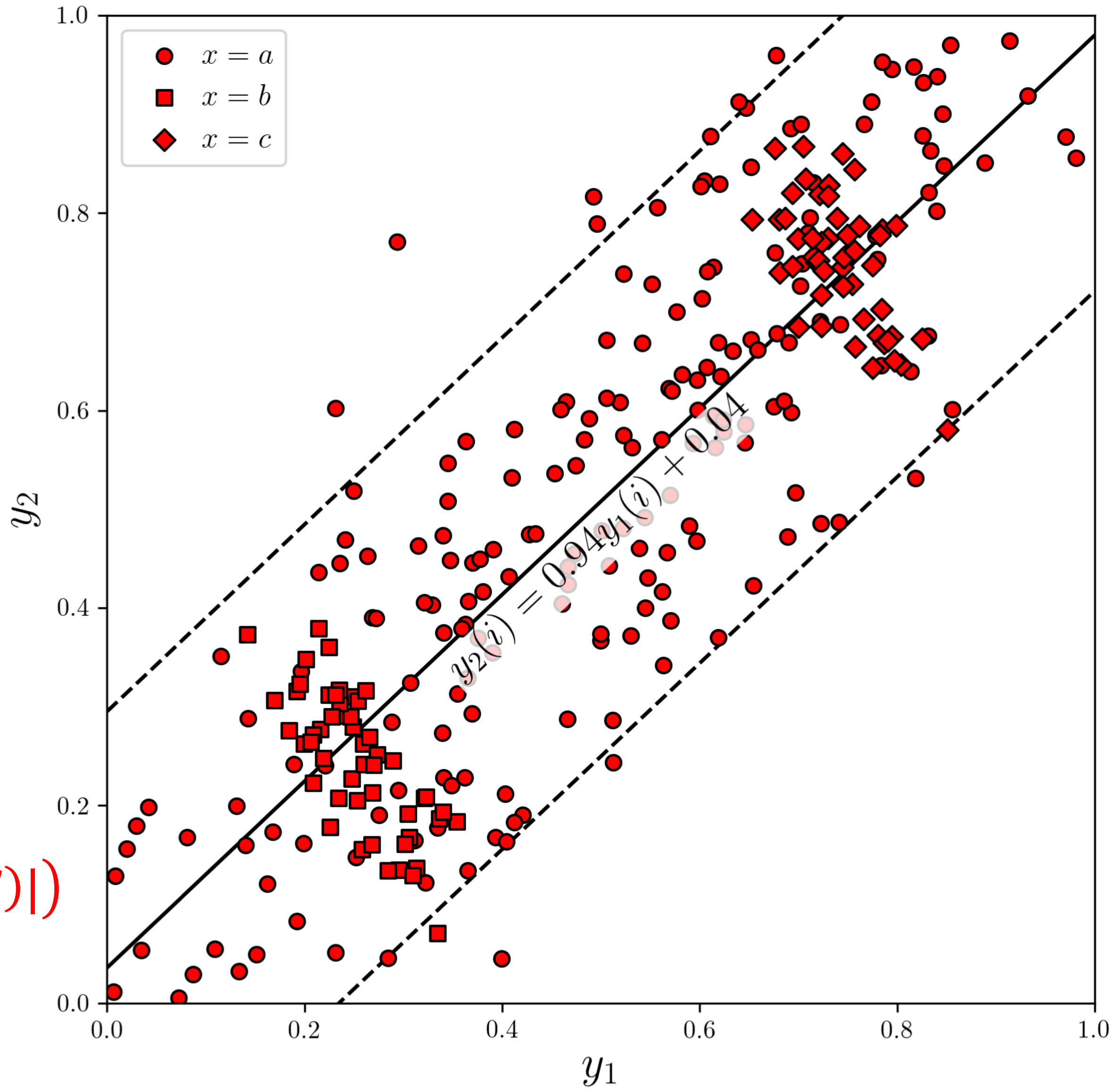
Selection language $\mathcal{L}_x = \{\sigma(i) = \pi_{j_1}(i) \wedge \cdots \wedge \pi_{j_l}(i)\}$

Optimize

$$f(Q) = \text{cov}(Q)^\gamma \text{eff}(Q)_+$$

with

- $Q = \{i \in S \colon \sigma(i) = \top\}$
- $\text{cov}(Q) = |Q|/|S|$
- $\text{eff}(Q) = \left(|r_{y_1,y_2}(Q)| - |r_{y_1,y_2}(S)|\right) / \left(1 - |r_{y_1,y_2}(S)|\right)$
- $r(Q) = \frac{1}{|Q|-1} \sum_{i \in Q} \left(\frac{\overline{y}_1(Q) - y_1(i)}{s_{y_1}(Q)}\right)\left(\frac{\overline{y}_2(Q) - y_2(i)}{s_{y_s}(Q)}\right)$



Subgroup discovery with multiple targets

Given

Sample $S \subseteq P$

Target variable $y_1, y_2 \colon P \to \mathbb{R}$

Features $x_j \colon P \to X_j$

Define

Propositions $\Pi_x = \{\pi_1, \ldots \pi_k\}$

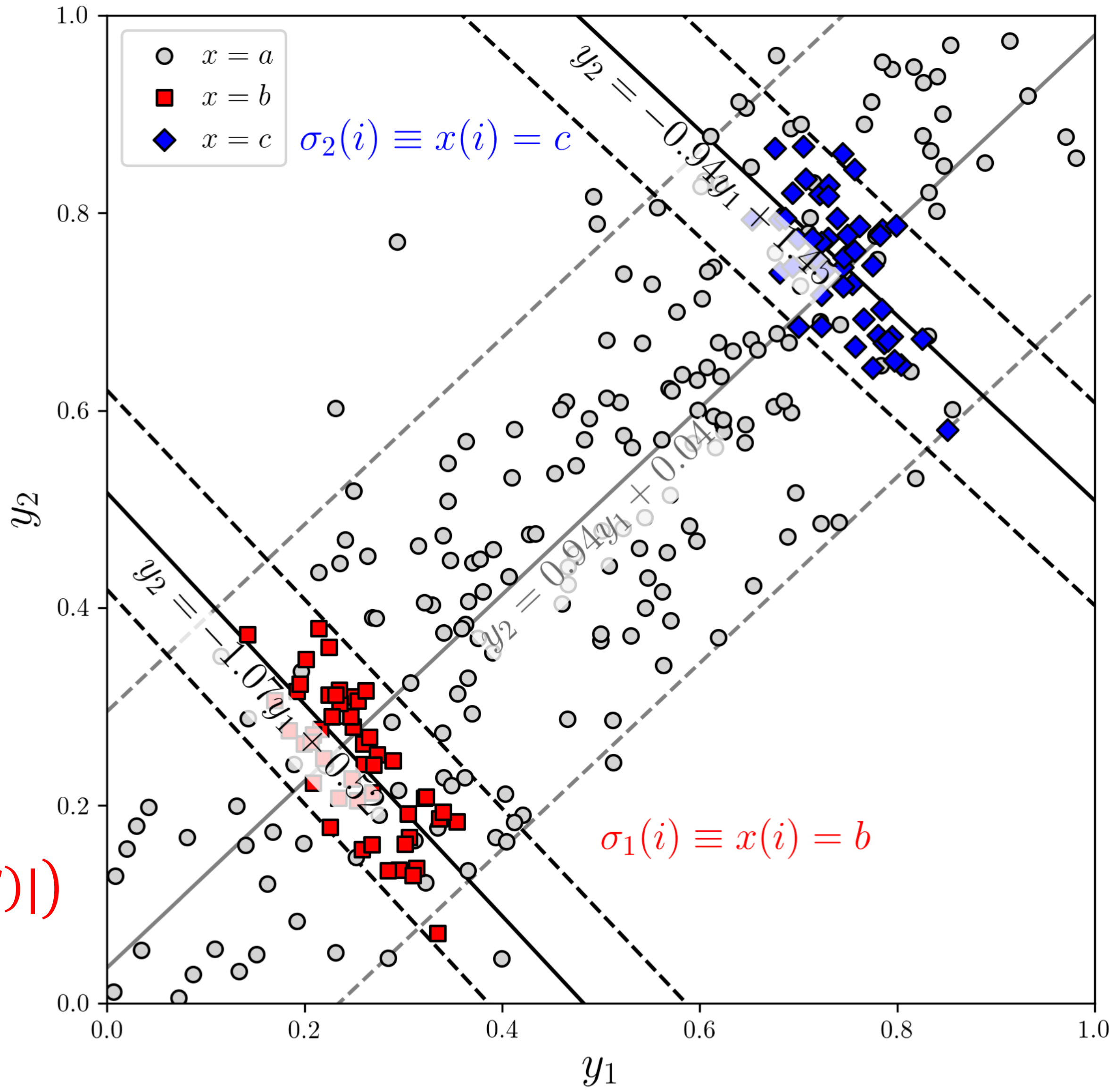
Selection language $\mathcal{L}_x = \{\sigma(i) = \pi_{j_1}(i) \wedge \cdots \wedge \pi_{j_l}(i)\}$

Optimize

$$f(Q) = \mathrm{cov}(Q)^\gamma \mathrm{eff}(Q)_+$$

with

- $Q = \{i \in S \colon \sigma(i) = \top\}$
- $\mathrm{cov}(Q) = |Q|/|S|$
- $\mathrm{eff}(Q) = \left(|r_{y_1,y_2}(Q)| - |r_{y_1,y_2}(S)|\right)/\left(1 - |r_{y_1,y_2}(S)|\right)$
- $r(Q) = \frac{1}{|Q|-1} \sum_{i \in Q} \left(\frac{\overline{y}_1(Q) - y_1(i)}{s_{y_1}(Q)}\right)\left(\frac{\overline{y}_2(Q) - y_2(i)}{s_{y_s}(Q)}\right)$



Application 2: Au structure/property relationship

Population

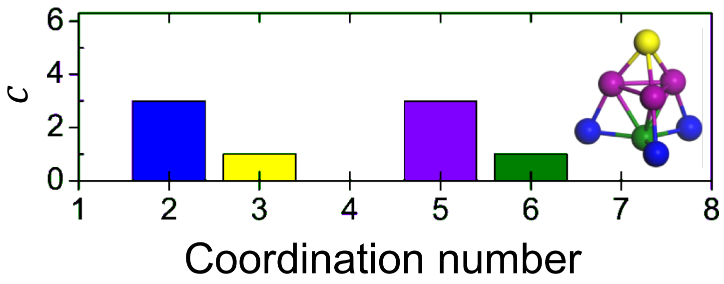
$$P = \{c : c \text{ conf. of Au5} - \text{Au14}\}$$

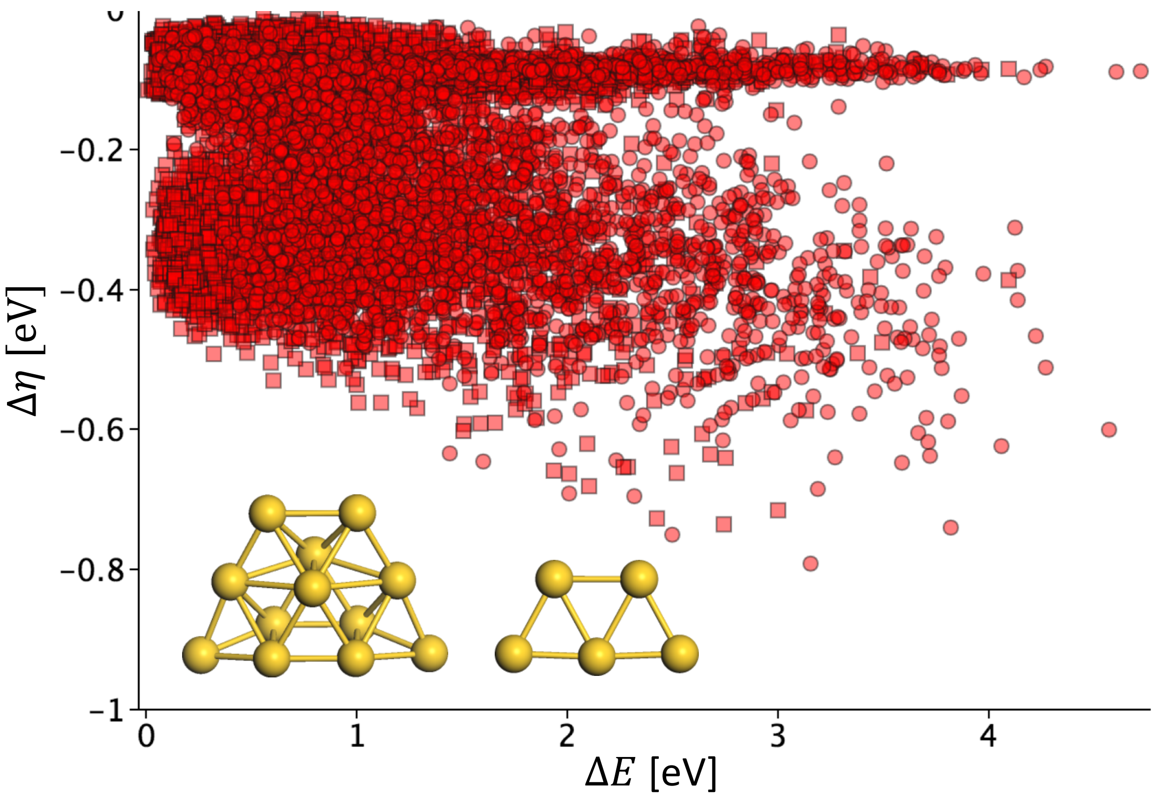
Targets

$$y_1 = \Delta E, y_2 = \Delta\eta \text{ chem. hardness}$$

Features

$$x \in \{a, c_1, c_2, c_3, c_4, c_5, c_6, r, \text{shape}, \text{Mo}_{\text{co}}, \text{Me}_{\text{co}}\}$$





Application 2: Au structure/property relationship

Population

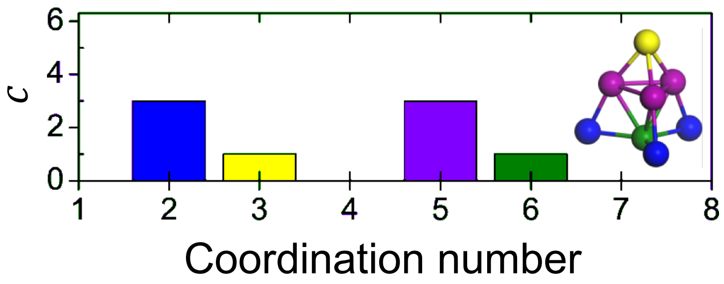
$$P = \{c : c \text{ conf. of Au5} - \text{Au14}\}$$

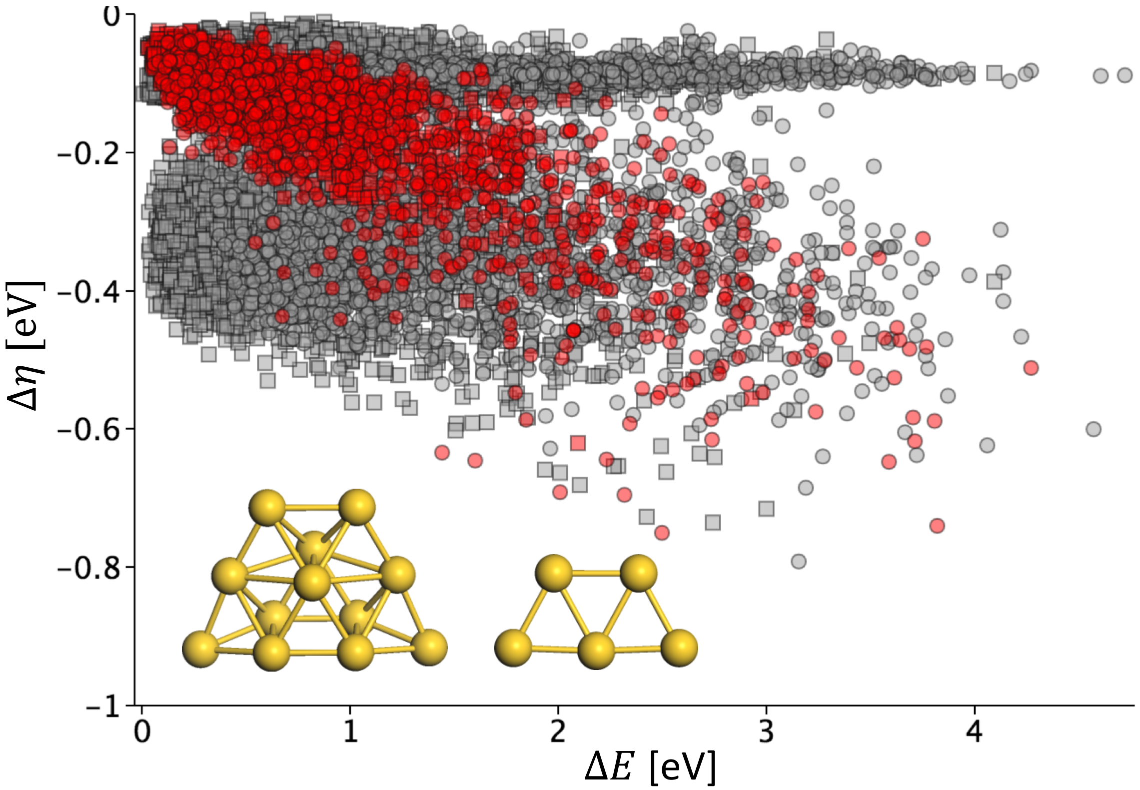
Targets

$$y_1 = \Delta E, y_2 = \Delta\eta \text{ chem. hardness}$$

Features

$$x \in \{a, c_1, c_2, c_3, c_4, c_5, c_6, r, \text{shape}, \text{Mo}_{\text{co}}, \text{Me}_{\text{co}}\}$$





Selector $\sigma(i) \equiv \text{even}\big(a(i)\big) \wedge (c_5(i) \leq 0.24) \wedge (\Delta\text{E}_{\text{vdw}}(i) \leq 0.18)$

Parameters $\text{cov}(\sigma) = 0.2$ $\text{eff}(\sigma) = 0.74$ $[r(Q) = -0.81, r(S) = -0.27]$

Summary

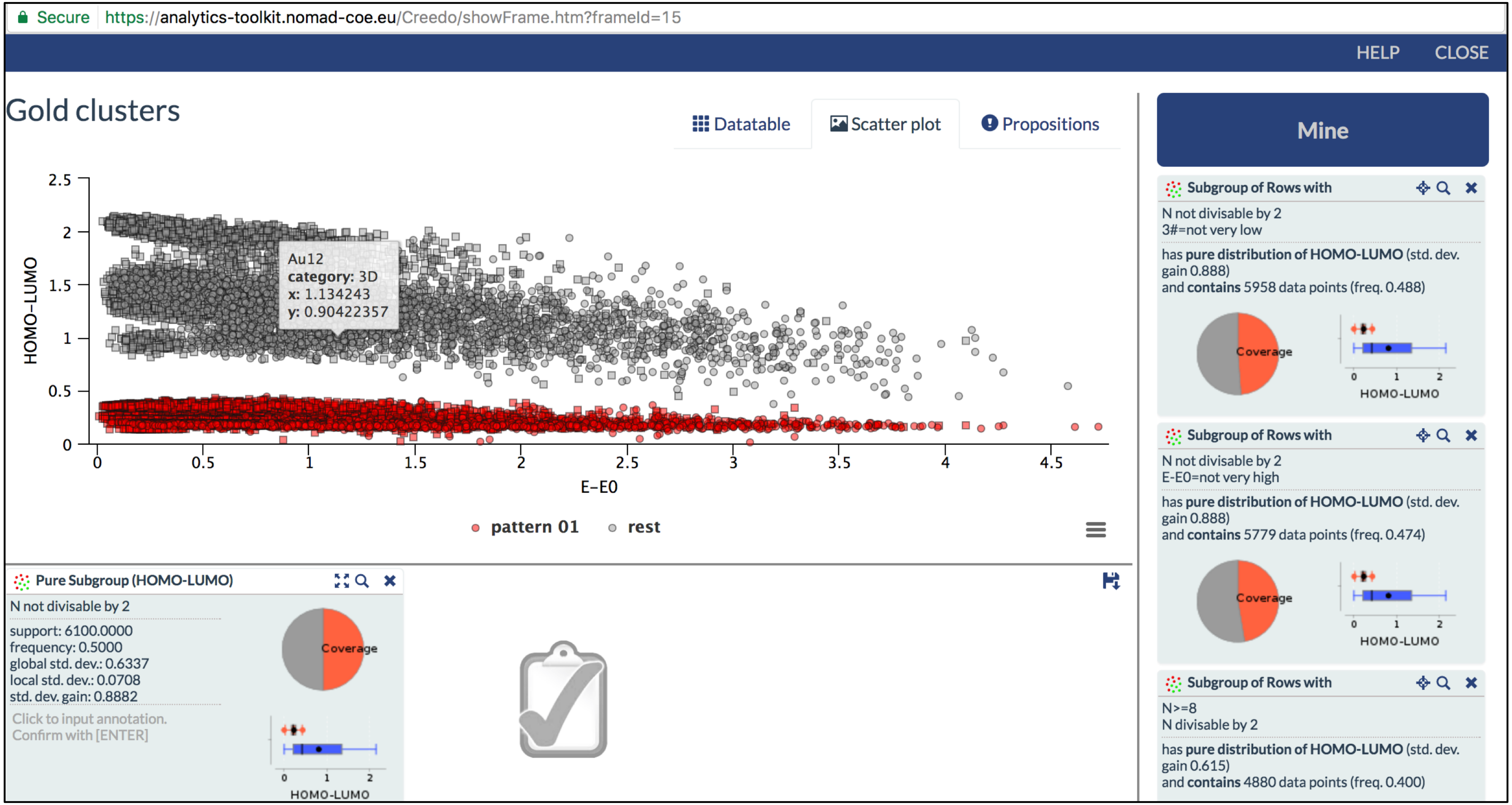
Topics

- Basic concepts: selectors, extensions, objective functions
- Application I: octet binary crystal structures
- A glimpse beyond: numeric and multiple targets
- Application II: Au structure/property relationship

References

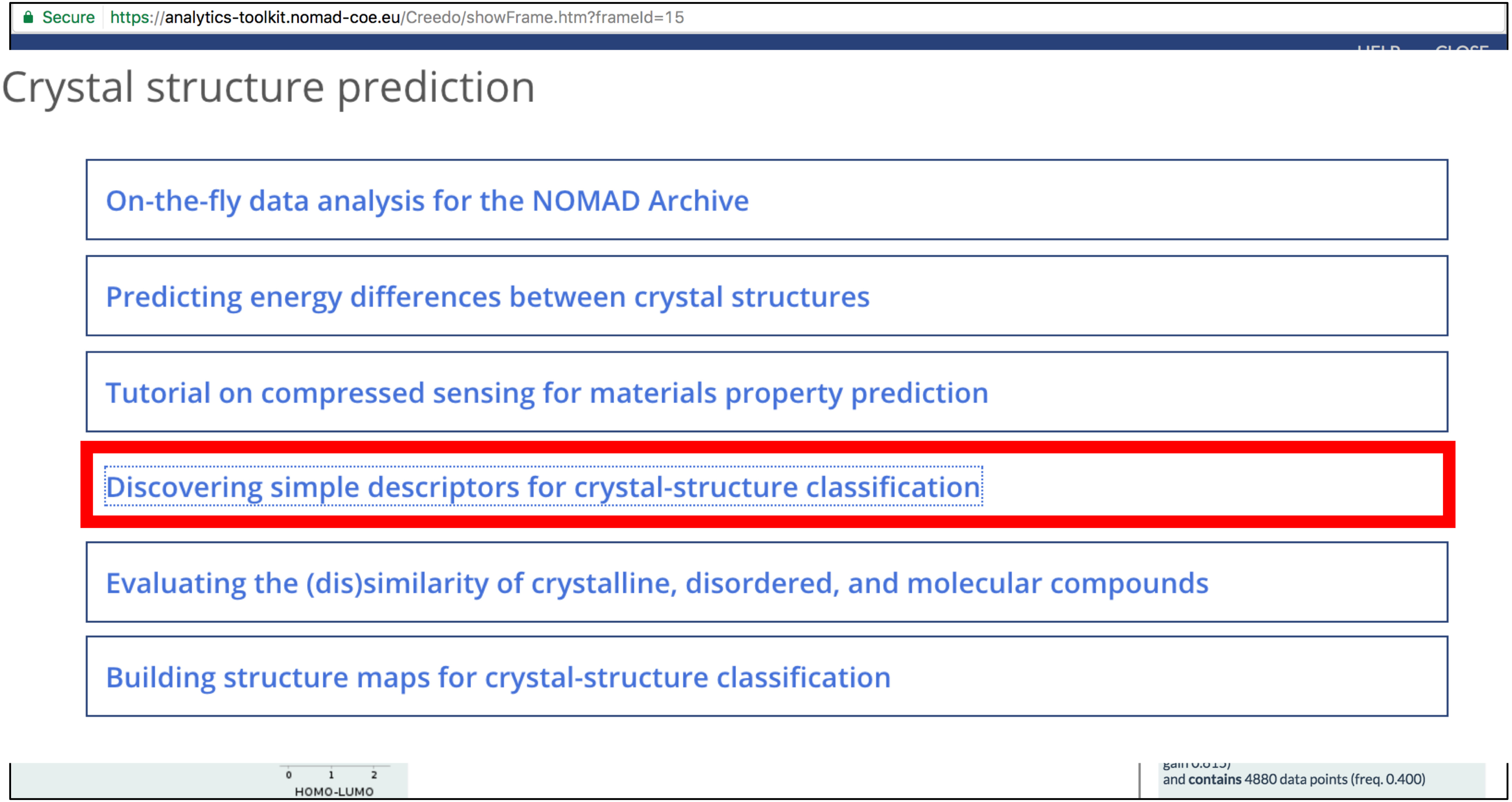
- **Boley, www.realkd.org: The power of saying 'I don't know'**
- Atzmueller, WIREs Data Mining Knowl Discov, 2015: Subgroup discovery – advanced review
- Friedman and Fisher, Stat Comput, 1999: Bump hunting in high-dimensional data
- **Goldsmith et al., New J Phys, 2017: Uncovering structure-property relationships by subgroup discovery**
- Boley et al., Data Min Knowl Disc, 2017: Identifying consistent statements about numerical data with dispersion-corrected subgroup discovery

Try it out for yourself



https://analytics-toolkit.nomad-coe.eu

Try it out for yourself



https://analytics-toolkit.nomad-coe.eu