NOMAD
NOVEL MATERIALS DISCOVERY

MAX-PLANCK-GESELLSCHAFT

HPC for Computationally and Data-Intensive Problems

Luca M. Ghiringhelli
Fritz-Haber-Institut der MPG, Berlin

SuperMUC-NG
Next-Gen Science Symposium
Munich, November 22, 2018

Data-driven materials science
The Big Picture

- Design of new materials:
  preparation, synthesis, and characterization is complex and costly

Data-driven materials science
The Big Picture

- Design of new materials:
  preparation, synthesis, and characterization is complex and costly
- About 240 000 *inorganic* materials are known to exist (Springer Materials)

NOVEL MATERIALS DISCOVERY

Data-driven materials science
The Big Picture

- Design of new materials:
  preparation, synthesis, and characterization is complex and costly
- About 240 000 *inorganic* materials are known to exist (Springer Materials)
- Basic properties determined for very few of them

Data-driven materials science
The Big Picture

- Design of new materials:
  preparation, synthesis, and characterization is complex and costly
- About 240 000 *inorganic* materials are known to exist (Springer Materials)
- Basic properties determined for very few of them
- Number of possible materials: practically infinite

Data-driven materials science
The Big Picture

- Design of new materials:
  preparation, synthesis, and characterization is complex and costly
- About 240 000 *inorganic* materials are known to exist (Springer Materials)
- Basic properties determined for very few of them
- Number of possible materials: practically infinite

⇨ New materials with superior properties exist but not yet known

Data-driven materials science
The Big Picture

- Design of new materials:
  preparation, synthesis, and characterization is complex and costly
- About 240 000 *inorganic* materials are known to exist (Springer Materials)
- Basic properties determined for very few of them
- Number of possible materials: practically infinite

⇨ New materials with superior properties exist but not yet known

- Data analytics tools will help to identify trends and anomalies in data and guide discovery of new materials

The NOMAD Laboratory
https://nomad-coe.eu

NOVEL MATERIALS DISCOVERY

The NOMAD Laboratory
A European Centre of Excellence

HOME PROJECT INDUSTRY TEAM RELATED PROJECTS NEWS PRESS KIT CONTACT US

Enter Search...

NOMAD REPOSITORY | THE ARCHIVE | ENCYCLOPEDIA | BIG-DATA ANALYTICS | ADVANCED GRAPHICS | HPC INFRASTRUCTURE | OUTREACH

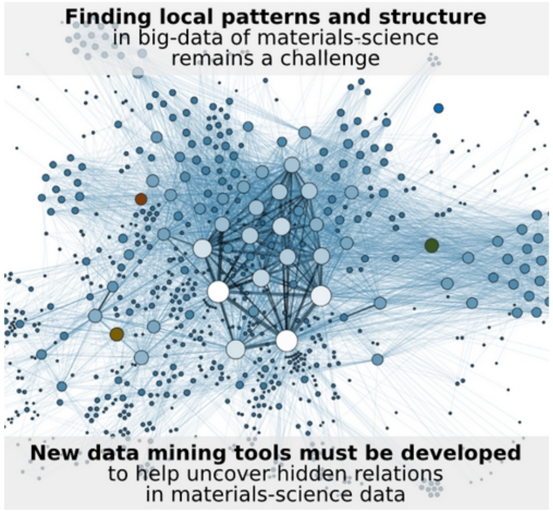
The Novel Materials Discovery (NOMAD) Laboratory maintains the largest Repository, for input and output files of all important computational materials science codes.

From its open-access data, it builds several ***Big-Data Services*** helping to advance materials science and engineering.

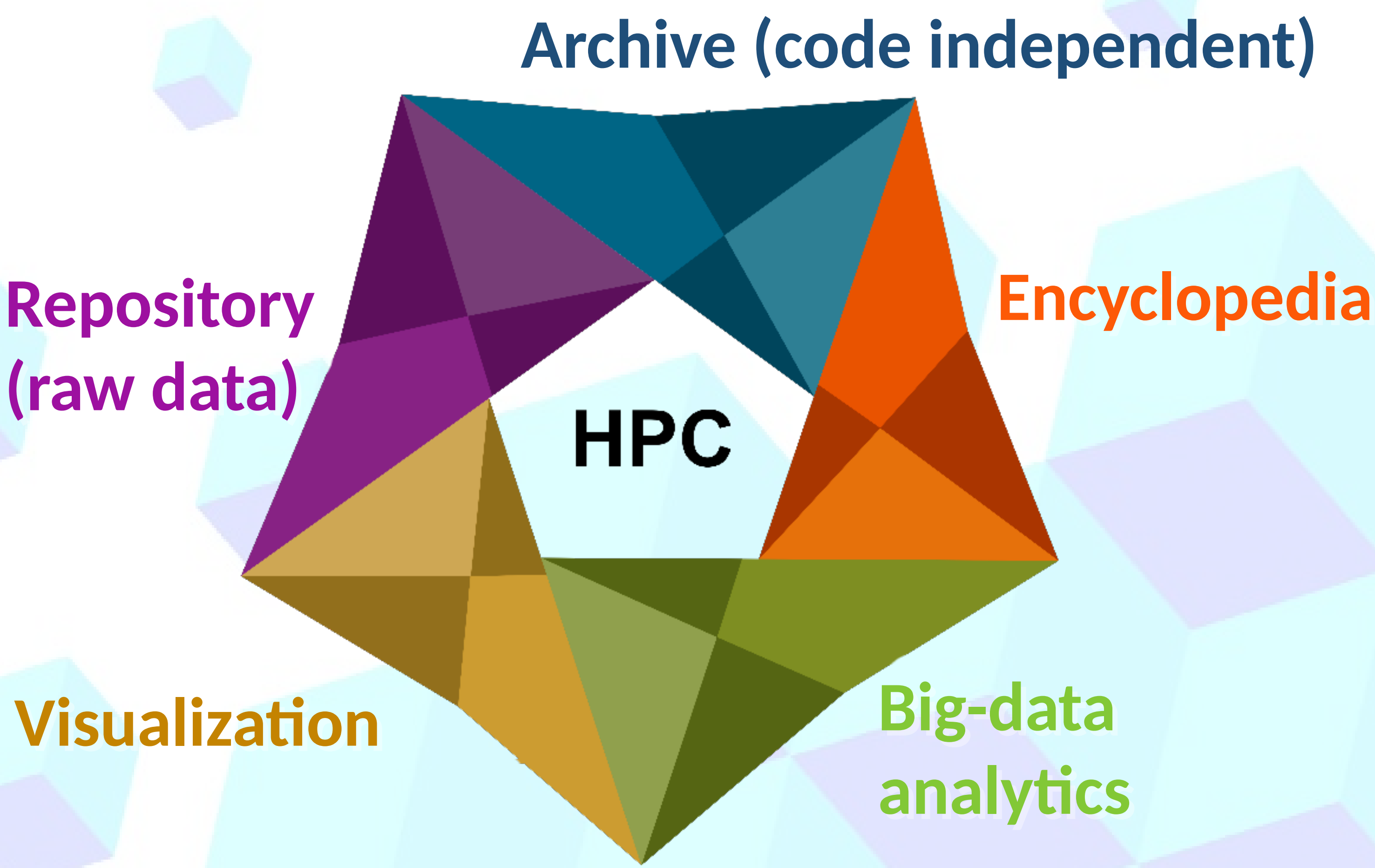
To learn more, click on the buttons above. You can also watch our 3-minute summary on the ***NOMAD Laboratory CoE*** at YouTube (or at YOUKU in China).

NOMAD Scope and Overview

NOMAD Success Stories



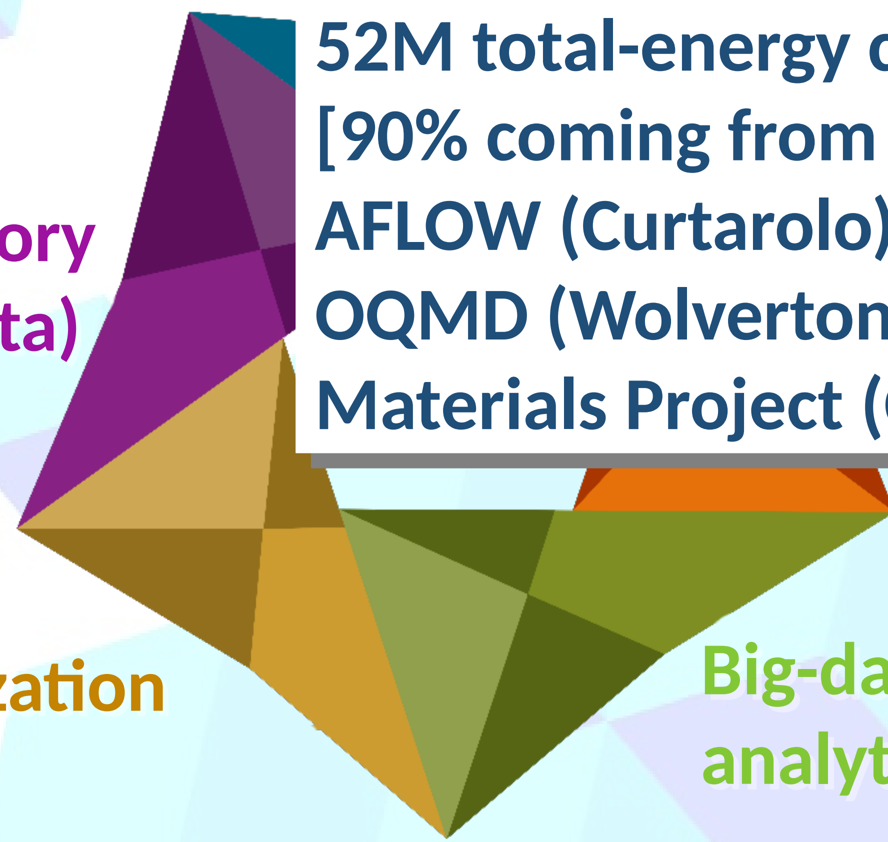
The NOMAD Laboratory
https://nomad-coe.eu



The NOMAD Laboratory
https://nomad-coe.eu

NOMAD NOVEL MATERIALS DISCOVERY

Archive (code independent)
52M total-energy calculations
[90% coming from
AFLOW (Curtarolo)
OQMD (Wolverton)
Materials Project (Ceder)]



Data analytics: an ideal flow chart

Training set
Calculate properties and functions
P_i, for many *materials*, i
E.g., Density-Functional Theory

Data analytics: an ideal flow chart

Training set
Calculate properties and functions P_i, for many *materials*, i
E.g., Density-Functional Theory

Descriptor
Find the appropriate descriptor $\boldsymbol{d}_i$, build a table:
$| \, i \, | \, \boldsymbol{d}_i \, | \, P_i \, |$

Data analytics: an ideal flow chart

Training set
Calculate properties and functions
P_i, for many *materials*, i
E.g., Density-Functional Theory

Descriptor
Find the appropriate
descriptor $\boldsymbol{d}_i$,
build a table:
$| \ i \ | \ \boldsymbol{d}_i \ | \ P_i \ |$

Learning
Find the function $P(\boldsymbol{d})$ for the table.
Build a **chart** for the property
Statistical learning

Data analytics: an ideal flow chart

Training set
Calculate properties and functions P_i, for many *materials*, i
E.g., Density-Functional Theory

Descriptor
Find the appropriate descriptor $\boldsymbol{d}_i$, build a table:
$| \; i \; | \; \boldsymbol{d}_i \; | \; P_i \; |$

Learning
Find the function $P(\boldsymbol{d})$ for the table.
Build a **chart** for the property
Statistical learning

Fast Prediction
Calculate properties and functions for new values of $\boldsymbol{d}$ (new materials)

Data analytics: an ideal flow chart

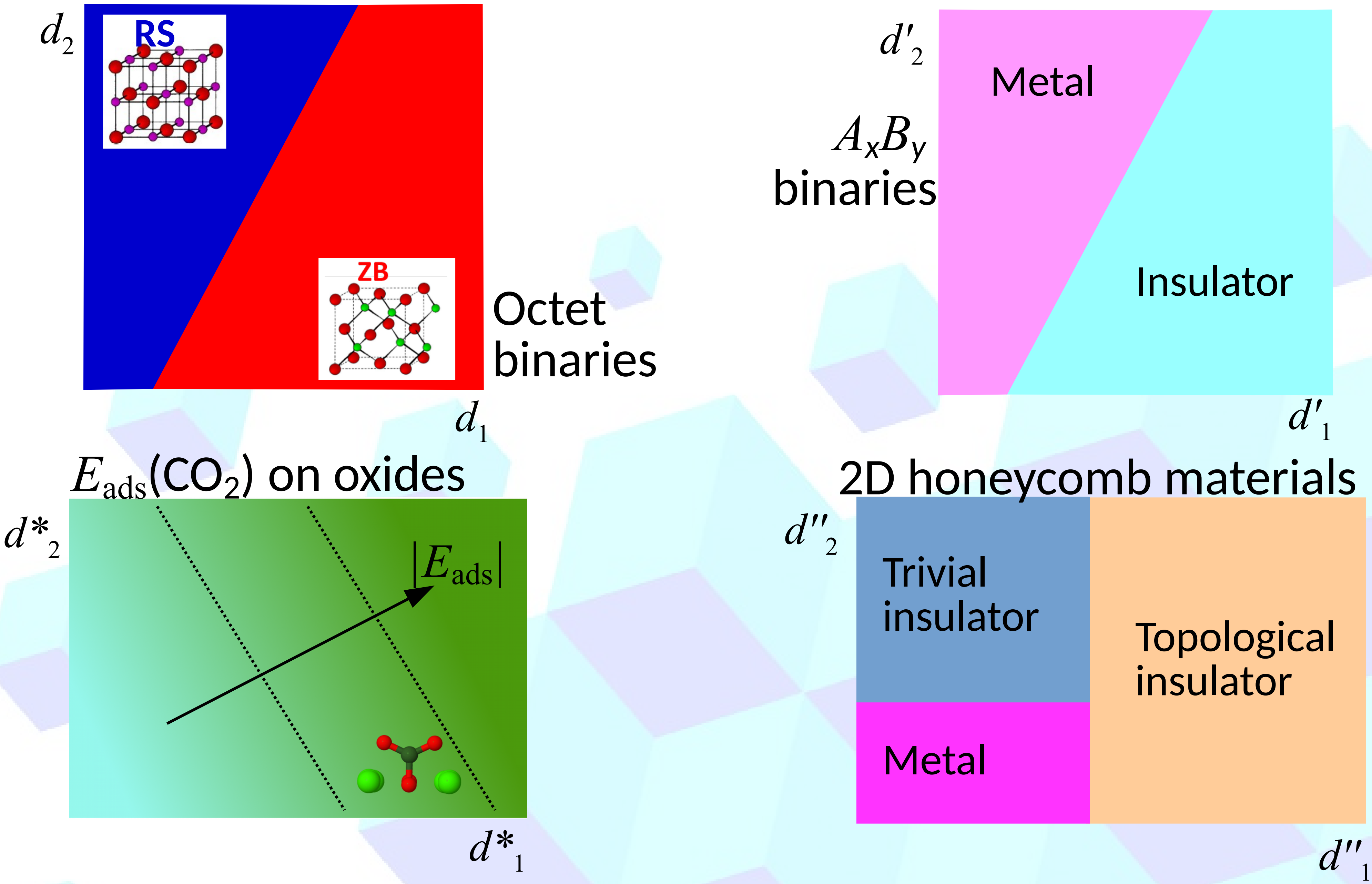
Training set
Calculate properties and functions
P_i, for many *materials*, i
E.g., Density-Functional Theory

Descriptor
Find the appropriate
descriptor $\boldsymbol{d}_i$,
build a table:
$| \; i \; | \; \boldsymbol{d}_i \; | \; P_i \; |$

Learning
Find the function $P(\boldsymbol{d})$ for the table.
Build a **chart** for the property
Statistical learning

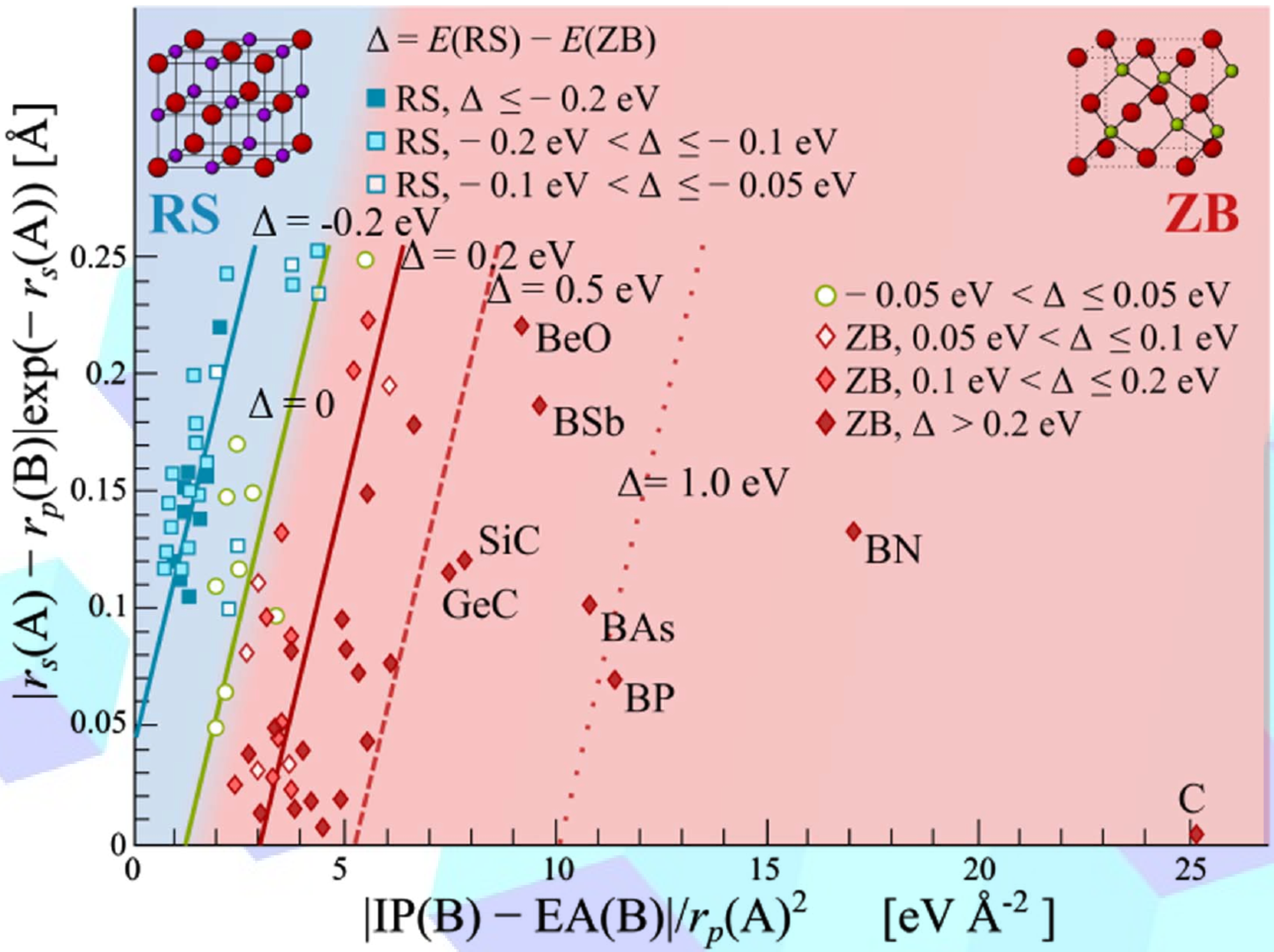
Fast Prediction
Calculate properties
and functions for
new values of $\boldsymbol{d}$
(new materials)

Learning/discovering maps of materials properties. A quantum many-body problem



Compressed sensing: the quest for descriptors and predictive models

Structure map with compressed-sensing algorithm, starting from 7 atomic features



L. Ghiringhelli *et al.*,
PRL 2015, NJP 2017

Data analytics with compressed sensing

Compressed sensing
Aim: finding descriptors and learning predictive models

Ansatz:

$$\boldsymbol{P} = c_1\boldsymbol{d}_1 + c_2\boldsymbol{d}_2 + \ldots + c_n\boldsymbol{d}_n$$

$\boldsymbol{P}$: property of interest
$\boldsymbol{d}_1, \ldots \boldsymbol{d}_n$: features, i.e., (nonlinear) functions of *primary features* (EA, IP, ...)
$c_1, \ldots c_\mathrm{n}$: unknown coefficients => as few as possible are *nonzero*

Data analytics with compressed sensing

Compressed sensing

Aim: finding descriptors and learning predictive models

Ansatz:

$\boldsymbol{P} = c_1\boldsymbol{d}_1 + c_2\boldsymbol{d}_2 + \ldots + c_n\boldsymbol{d}_n$

$\boldsymbol{P}$: property of interest

$\boldsymbol{d}_1, \ldots \boldsymbol{d}_n$: features, i.e., (nonlinear) functions of *primary features* (EA, IP, ...)

$c_1, \ldots c_n$: unknown coefficients => as few as possible are *nonzero*

d_i: iterative construction with features and operators (+, ×, /, ², ...)

$$\frac{\mathrm{IP(B)} - \mathrm{EA(B)}}{r_p(\mathrm{A})^2}, \quad \frac{|r_s(\mathrm{A}) - r_p(\mathrm{B})|}{\exp(r_s(\mathrm{A}))}$$

Data analytics with compressed sensing

Compressed sensing
Aim: finding descriptors and learning predictive models

Ansatz:
$$\boldsymbol{P} = c_1\boldsymbol{d}_1 + c_2\boldsymbol{d}_2 + \ldots + c_n\boldsymbol{d}_n$$

$\boldsymbol{P}$: property of interest
$\boldsymbol{d}_1, \ldots \boldsymbol{d}_n$: features, i.e., (nonlinear) functions of *primary features* (EA, IP, ...)
$c_1, \ldots c_n$: unknown coefficients => as few as possible are *nonzero*

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_0$$

***d_i*: iterative construction with features and operators (+, ×, /, ², ...)**

$$\frac{\mathrm{IP(B)} - \mathrm{EA(B)}}{r_p(\mathrm{A})^2}, \quad \frac{|r_s(\mathrm{A}) - r_p(\mathrm{B})|}{\exp(r_s(\mathrm{A}))}$$

SISSO: sure independence screening plus sparsifying operator

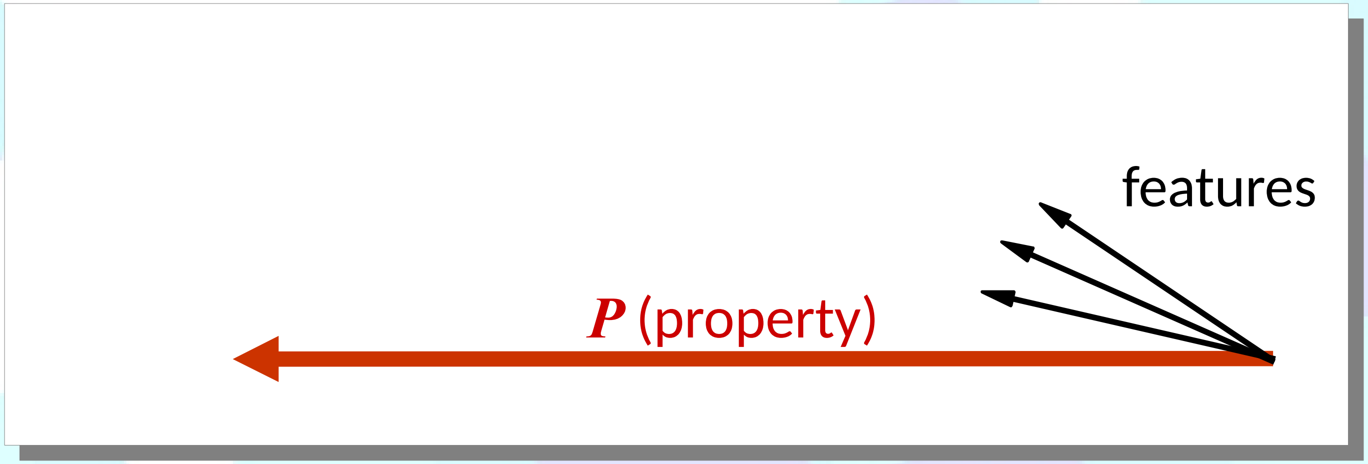
$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_0$$

Huge "feature space" (e.g., 100 billion elements)
constructed from primary features (e.g., 15 - 25 elements)

SISSO: sure independence screening plus sparsifying operator

$$\underset{\boldsymbol{c} \in \mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2 + \lambda \|\boldsymbol{c}\|_0$$

Huge “feature space” (e.g., 100 billion elements)
constructed from primary features (e.g., 15 - 25 elements)

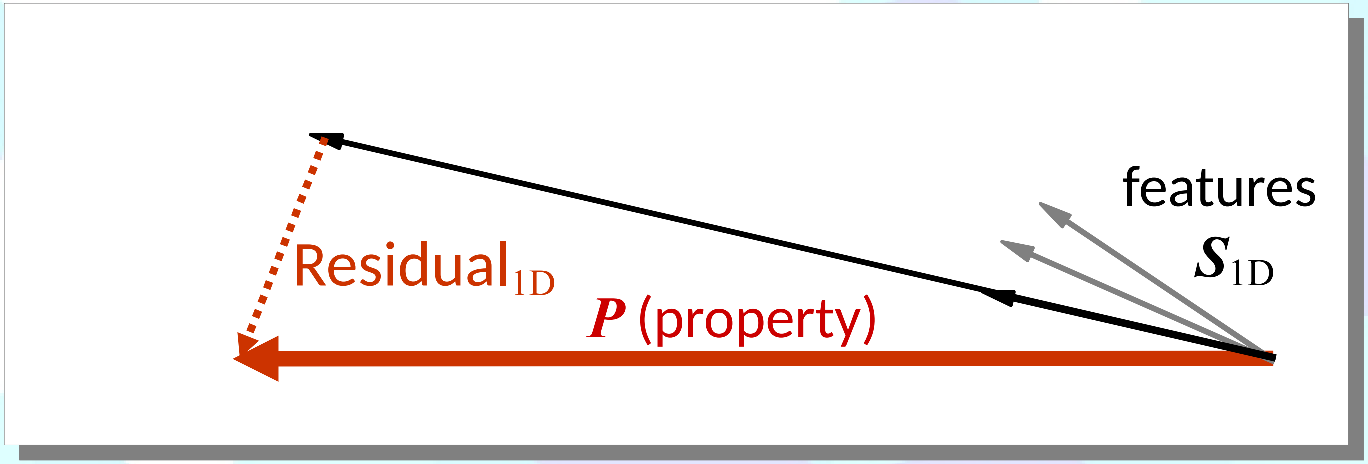


Ouyang, Curtarolo, Ahmetcik, Scheffler & LMG, PRM (2018) arXiv:1710.03319

SISSO: sure independence screening plus sparsifying operator

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P}-\boldsymbol{D}\boldsymbol{c}\|_2^2+\lambda\|\boldsymbol{c}\|_0$$

Huge "feature space" (e.g., 100 billion elements)
constructed from primary features (e.g., 15 - 25 elements)

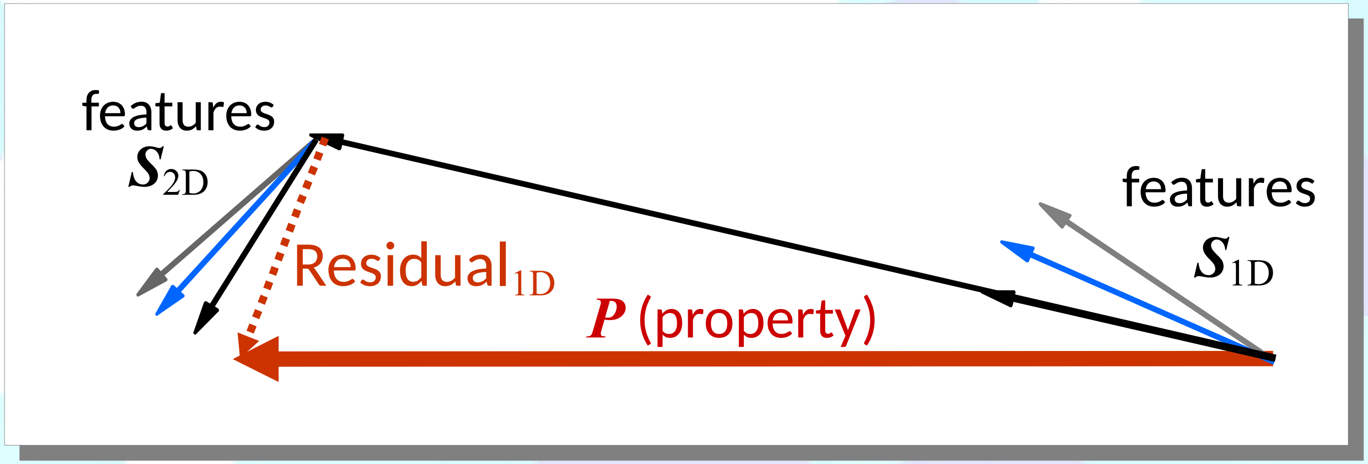


Ouyang, Curtarolo, Ahmetcik, Scheffler & LMG, PRM (2018) arXiv:1710.03319

SISSO: sure independence screening plus sparsifying operator

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_0$$

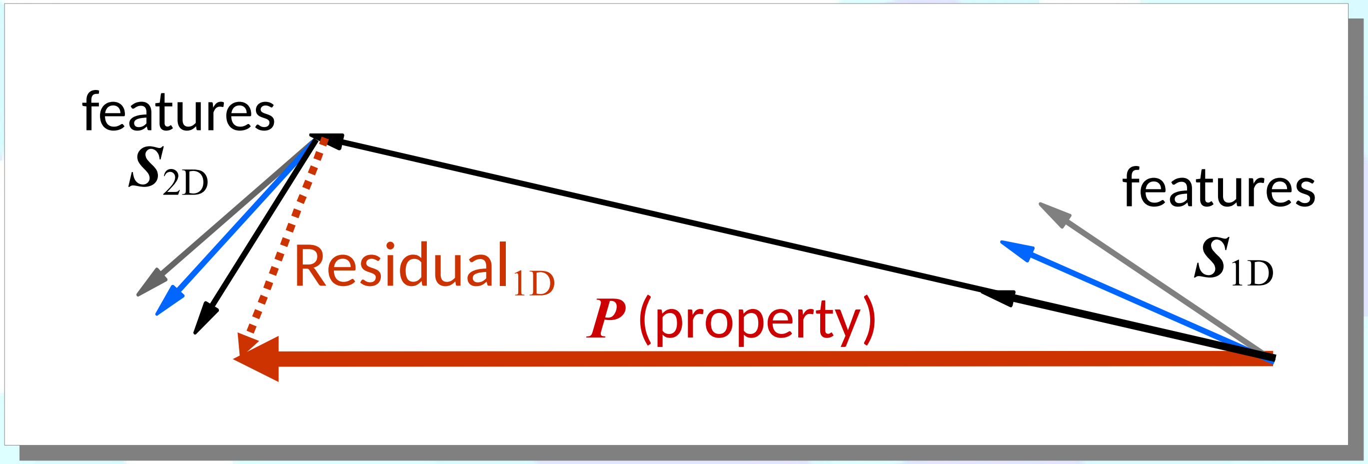
Huge “feature space” (e.g., 100 billion elements)
constructed from primary features (e.g., 15 - 25 elements)



Ouyang, Curtarolo, Ahmetcik, Scheffler & LMG, PRM (2018) arXiv:1710.03319

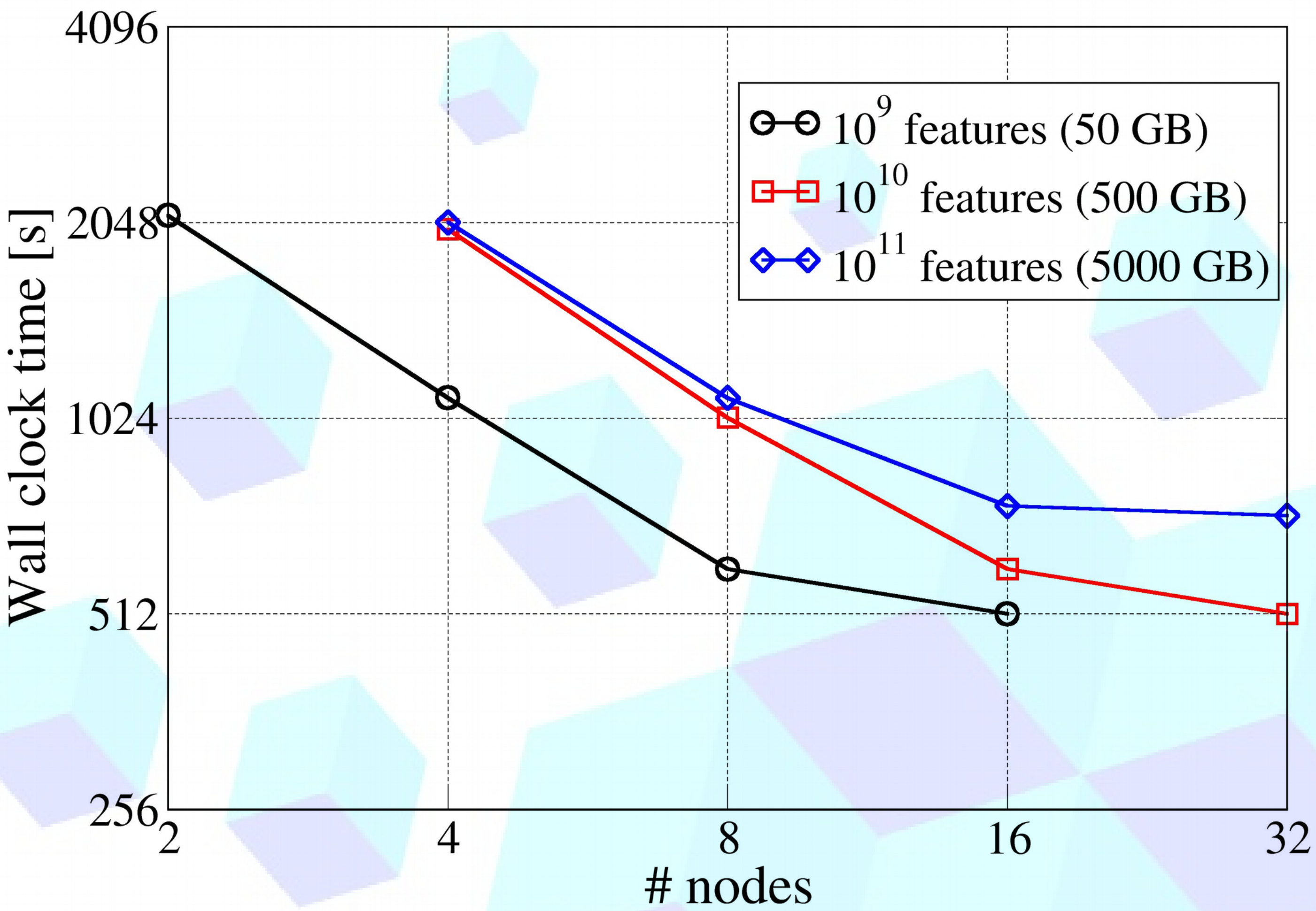
SISSO: sure independence screening plus sparsifying operator

Embarrassingly parallel
+ SIS: independent scalar products of features on property or residual)
- partial ranking
+ SO: independent least square regression
- partial ranking
+ outer parallelization for cross validation
- smart strategies needed for matrix storage



Ouyang, Curtarolo, Ahmetcik, Scheffler & LMG, PRM (2018) arXiv:1710.03319

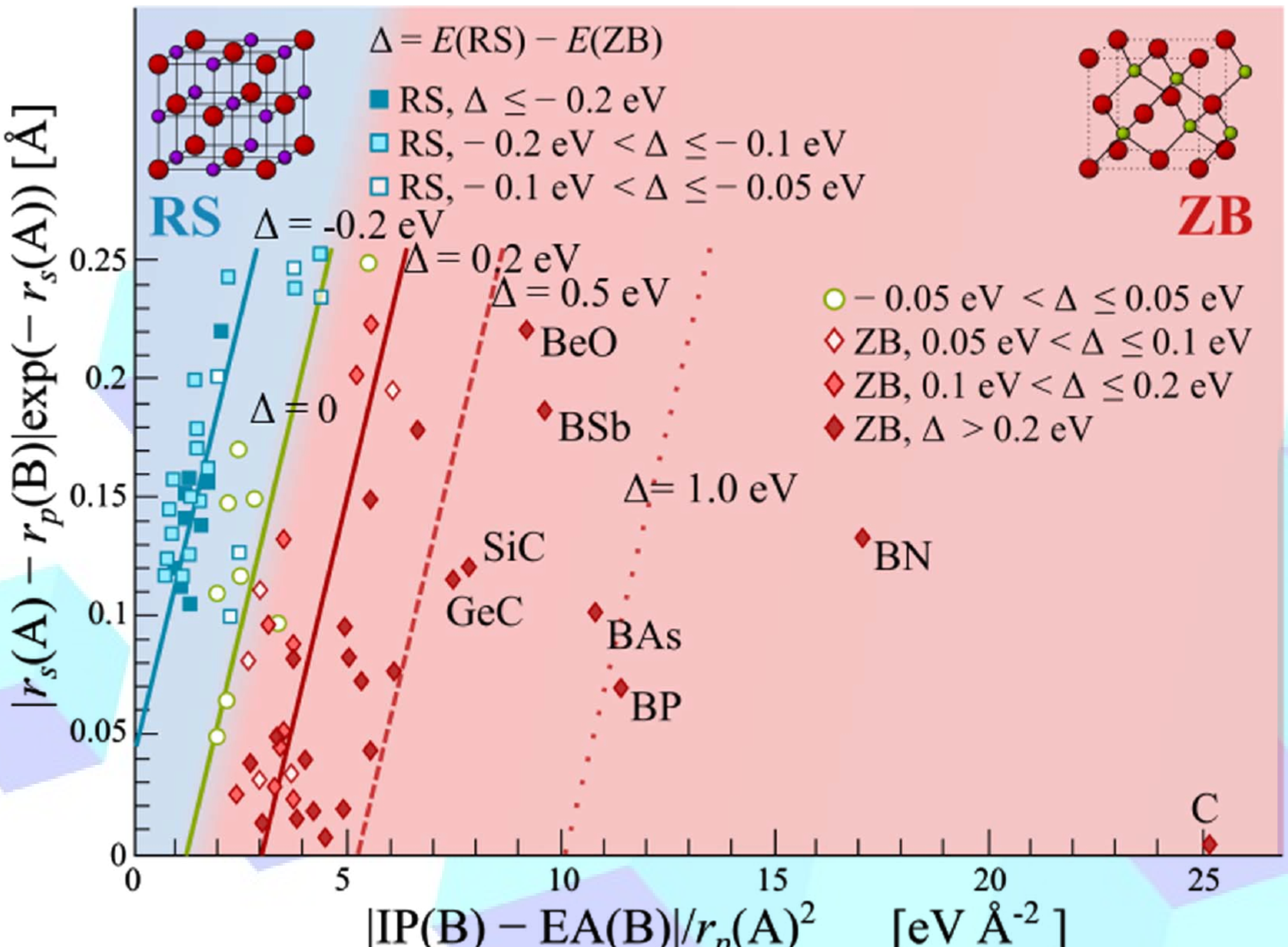
SISSO: sure independence screening plus sparsifying operator



Ouyang, Curtarolo, Ahmetcik, Scheffler & LMG, PRM (2018) arXiv:1710.03319

Compressed sensing: the quest for descriptors and predictive models

Structure map with compressed-sensing algorithm, starting from 7 atomic features



L. Ghiringhelli *et al.*,
PRL 2015, NJP 2017

Charts/maps of materials

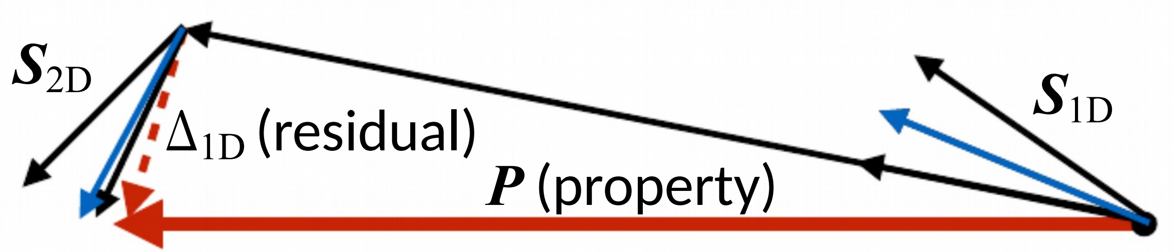
$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}}(\underbrace{\|\boldsymbol{P}-\boldsymbol{D}\boldsymbol{c}\|_2^2} + \lambda\|\boldsymbol{c}\|_0)$$

New cost function to be minimized: overlap of *convex* domains

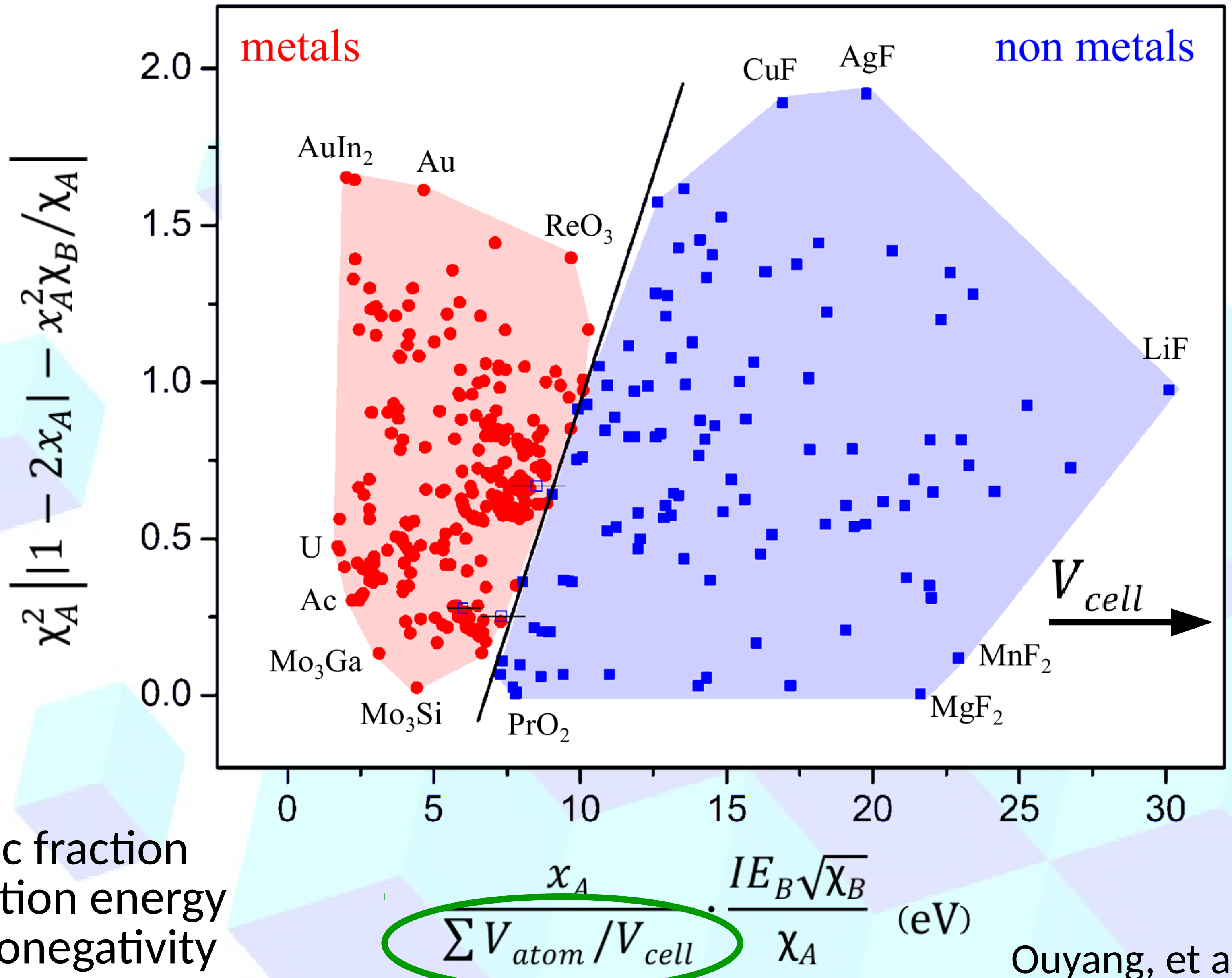
$$\hat{\boldsymbol{c}} \equiv \arg\min_{\boldsymbol{c}} \left(\sum_{i=1}^{M-1} \sum_{j=i+1}^{M} O_{ij} + \lambda\|\boldsymbol{c}\|_0 \right)$$

Number of data points in the overlap region, as function of selected ***d***

Iterative generation of feature subspaces



SISSO: metal/nonmetal classification of binary materials



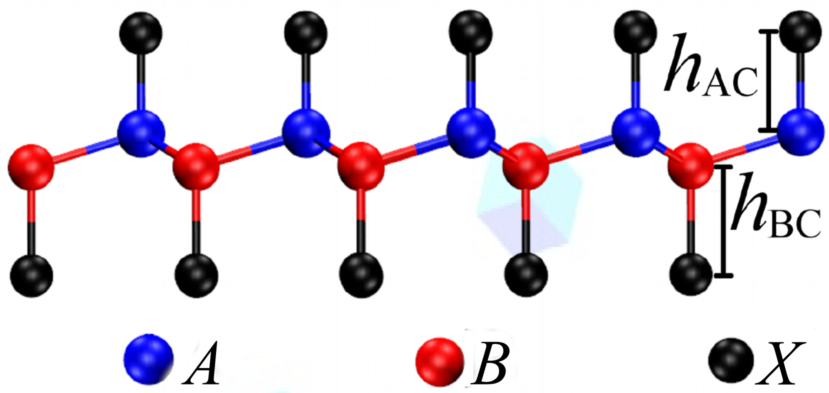
x Atomic fraction

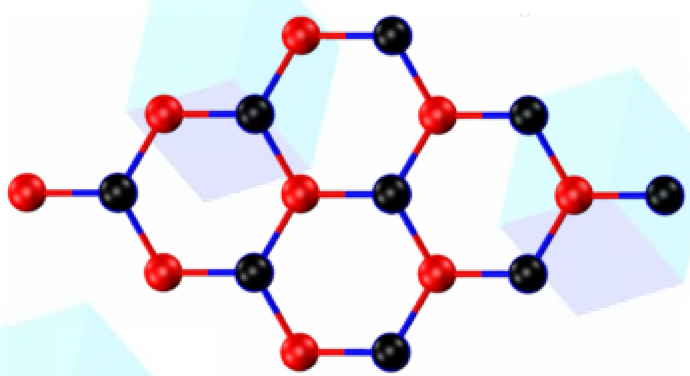
IE Ionization energy

χ Electronegativity

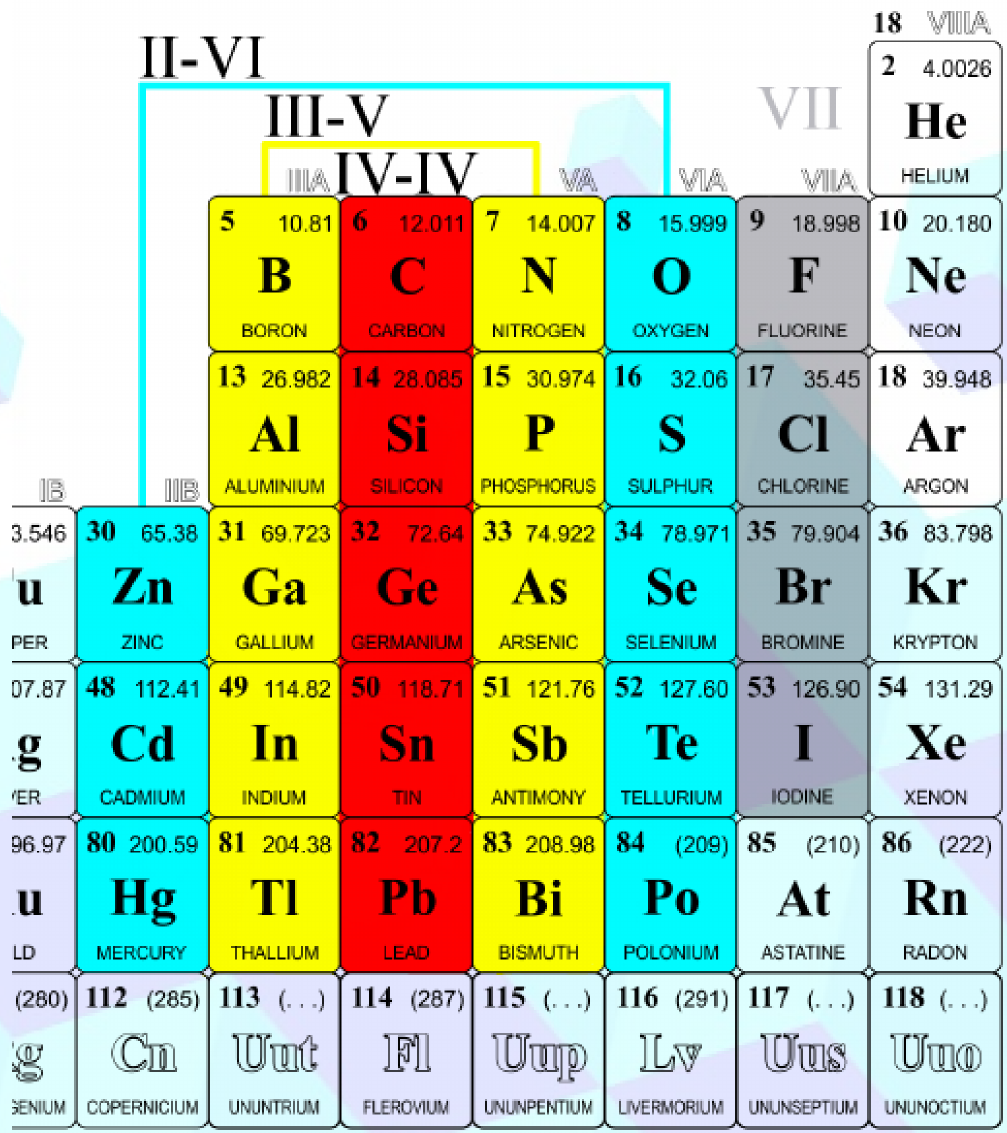
Ouyang, et al. PRM (2018)

SISSO: predicting novel honeycomb (~2D) topological insulators



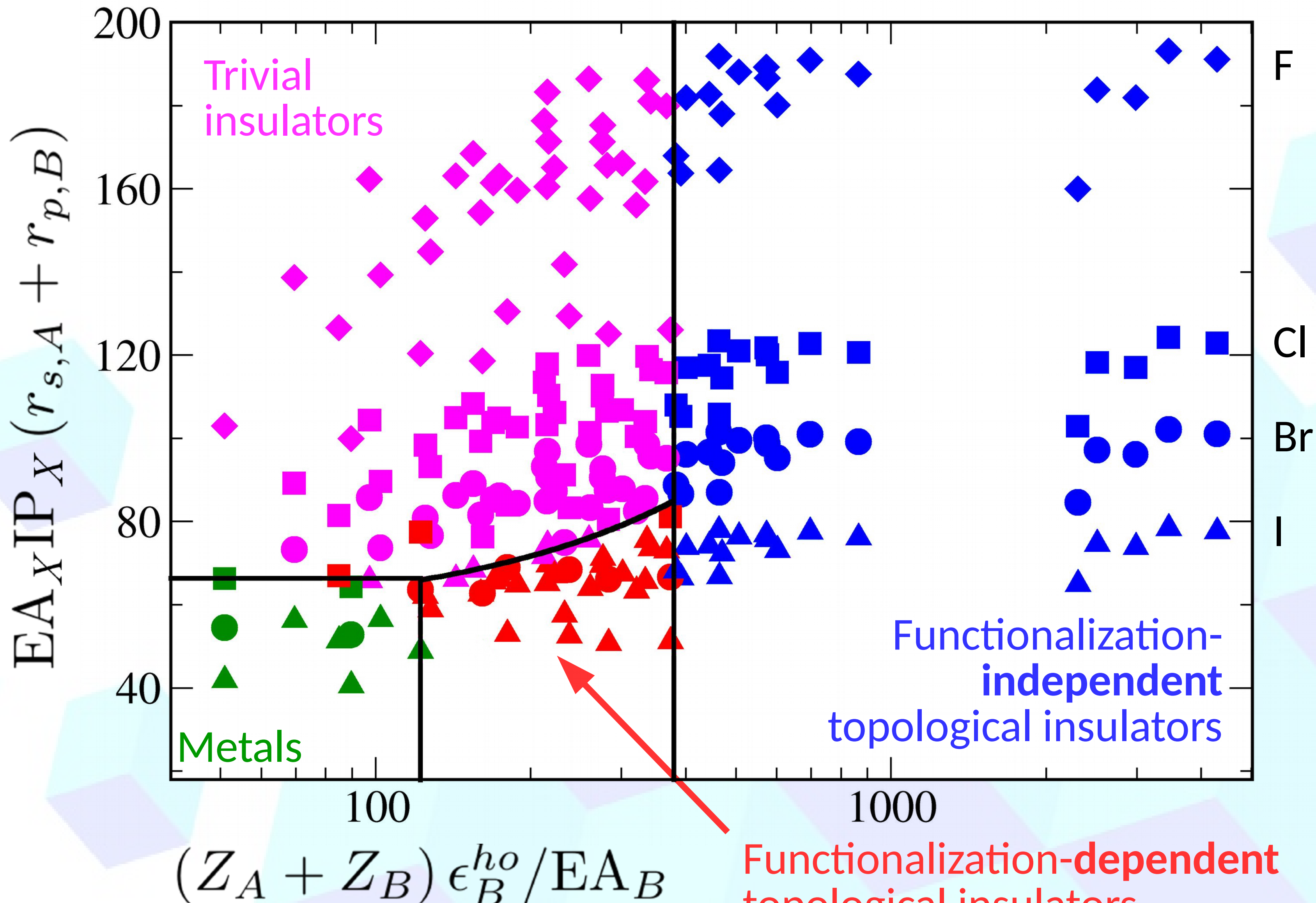


A	II	III	IV
B	VI	V	IV
X	VII	VII	VII



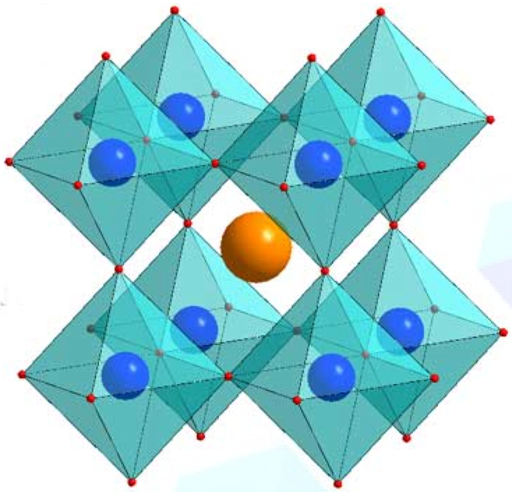
Data source: high throughput DFT (FHI-aims, Carlos Mera Acosta)

SISSO: predicting novel honeycomb (~2D) topological insulators



Functionalization-**dependent** topological insulators

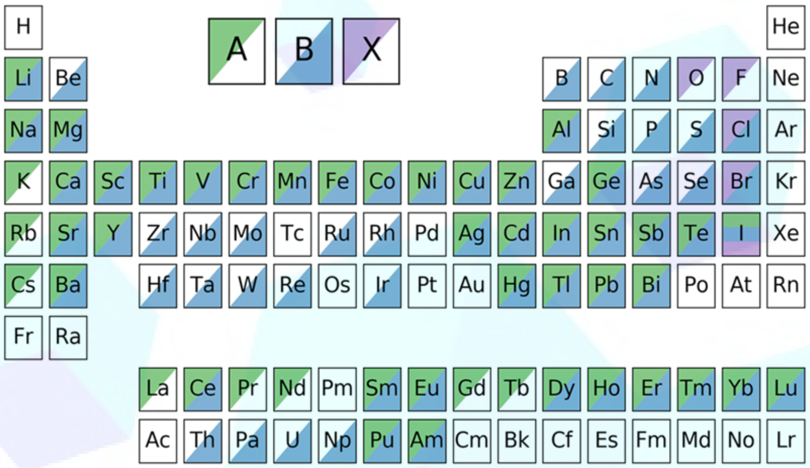
Perovskites' stability:
Improving on Goldschmidt Tolerance Factor



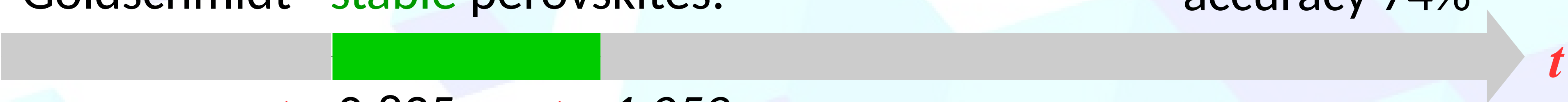
ABX$_3$

$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)}$$

Ionic radius



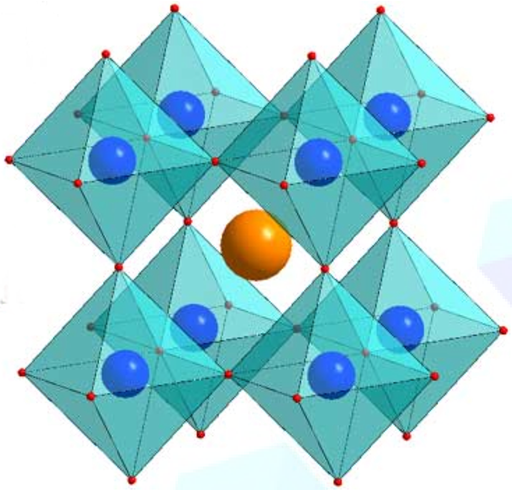
Goldschmidt* stable perovskites: accuracy 74%



t = 0.825 t = 1.059

Bartel, Sutton, Goldsmith, Ouyang, Musgrave, LMG &Scheffler, arXiv:1801.07700 (2018)

Perovskites' stability: Improving on Goldschmidt Tolerance Factor



ABX₃

$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)}$$

Ionic radius

Oxidation state

$$\tau = \boxed{\frac{r_X}{r_B}} - n_A\left(n_A - \frac{r_A/r_B}{\ln(r_A/r_B)}\right)$$

1 / μ = Octahedral factor

Goldschmidt* stable perovskites: accuracy 74%

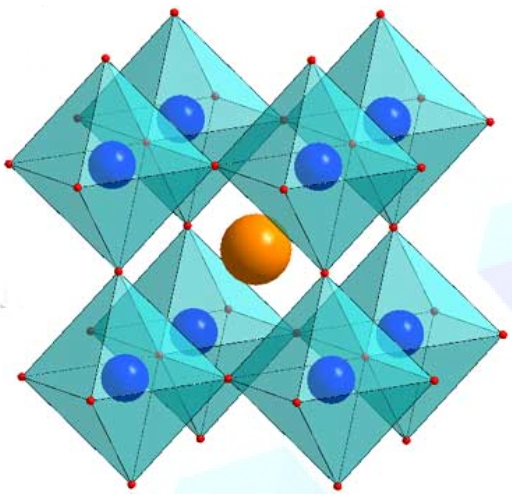
t = 0.825 t = 1.059 t

Our stable perovskites: accuracy 92%

τ = 4.18 τ

Bartel, Sutton, Goldsmith, Ouyang, Musgrave, LMG &Scheffler, arXiv:1801.07700 (2018)

Perovskites' stability: Improving on Goldschmidt Tolerance Factor



ABX_3

$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)}$$

Ionic radius

$$\tau = \boxed{\frac{r_X}{r_B}} - n_A\left(n_A - \frac{r_A/r_B}{\ln(r_A/r_B)}\right)$$

Oxidation state

$1/\mu$ = Octahedral factor

Goldschmidt* stable perovskites: accuracy 74%

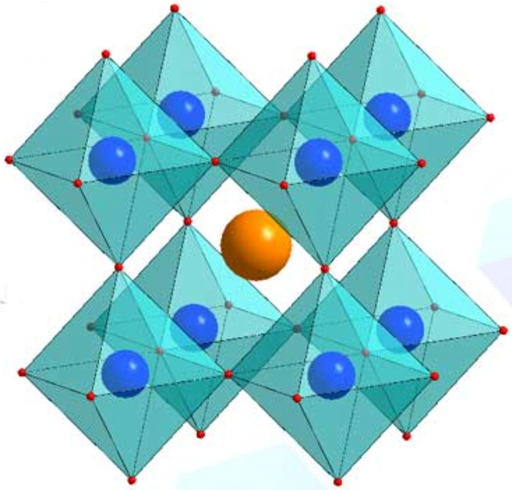
t = 0.825 t = 1.059 t

Our stable perovskites: accuracy 99%

τ = 3.31 τ = 4.18 τ = 5.92 τ

Bartel, Sutton, Goldsmith, Ouyang, Musgrave, LMG &Scheffler, arXiv:1801.07700 (2018)

Perovskites' stability:
Improving on Goldschmidt Tolerance Factor



ABX₃

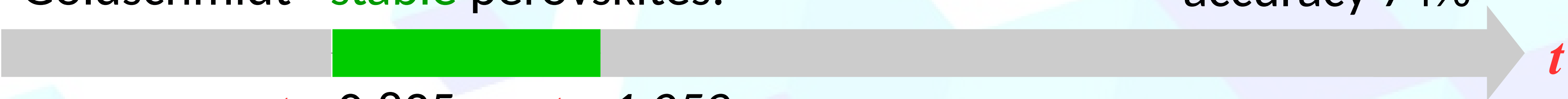
$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)}$$ → Ionic radius

$$\tau = \boxed{\frac{r_X}{r_B}} - n_A\left(n_A - \frac{r_A/r_B}{\ln(r_A/r_B)}\right)$$

→ Oxidation state

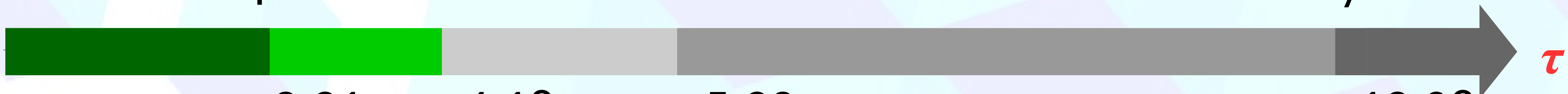
→ 1 / μ = Octahedral factor

Goldschmidt* stable perovskites: accuracy 74%



t = 0.825 t = 1.059

Our stable perovskites: accuracy 100%



τ = 3.31 τ = 4.18 τ = 5.92 τ = 12.08

Bartel, Sutton, Goldsmith, Ouyang, Musgrave, LMG &Scheffler, arXiv:1801.07700 (2018)

… and more

Continuous property

- Adsorption energy of O on metal-oxide surfaces
- Adsorption energy and OCO angle of adsorbed CO_2 on metal-oxide surfaces
- Adsorption energy of metal atoms on metal-alloys surfaces

Features: atoms (of the surface) and pristine surface

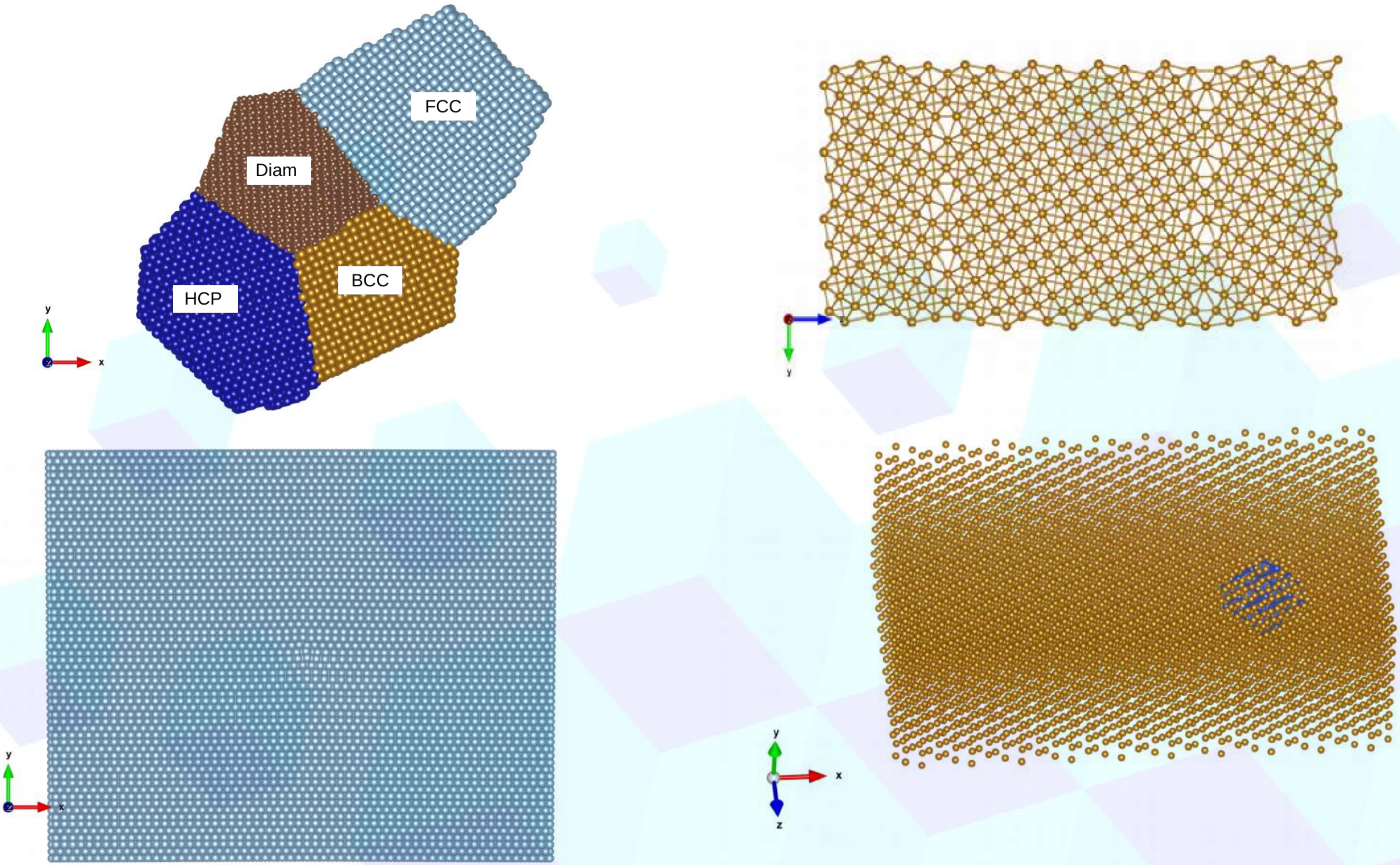
Classification

- Tetradymite 5-component 3d topological insulators (vs trivial insulators) arXiv:1808.04733

Features: atoms

… and further more

Convolutional neural networks for (local) crystal-structure recognition



Acknowledgements

NOVEL MATERIALS DISCOVERY

https://nomad-coe.eu

SISSO and metal/insulator proof of concepts

Runhai Ouyang, Emre Ahmetcik, Stefano Curtarolo

Topological Insulators

Carlos Mera Acosta, Adalberto Fazzio, Runhai Ouyang, Christian Carbogno

Perovskites

Christopher J. Bartel, Christopher Sutton, Bryan R. Goldsmith, Runhai Ouyang, Charles B. Musgrave

Convolutional neural networks for structure recognition

Angelo Ziletti, Andreas Leitherer, Devinder Kumar

NOMAD PIs

Angel Rubio (MPSD Hamburg), Risto Nieminen (Aalto Univ. Helsinki), Francesc Illas (Univ. Barcelona), Daan Frenkel (Univ. Cambridge), Claudia Draxl (HU Berlin), Alessandro De Vita (Kings College London), Kristian Thygesen (DTU Lyngby); Kimmo Koski (CSC Helsinki), Stefan Heinzel (MPSCD Garching), Jose Maria Cela (BSC Barcelona), Dieter Kranzlmüller (LRZ Munich); Ciaran Clissman (Pintail Dublin)

All the above

Matthias Scheffler

BBDC BERLIN BIG DATA CENTER

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