BARCELONA 2019

NOMAD

NOVEL MATERIALS DISCOVERY

FAIRmat

MAX-PLANCK-GESELLSCHAFT

(Big-)data analytics for materials science

Luca M. Ghiringhelli

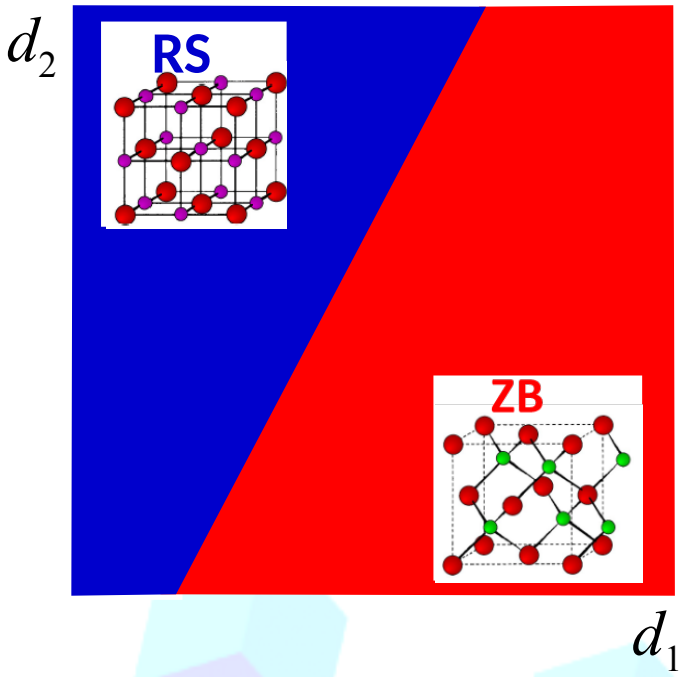
FRITZ-HABER-INSTITUT

MAX-PLANCK-GESELLSCHAFT

Hands-On DFT and Beyond:
High-throughput screening and big-data analytics,
towards exascale computational materials science
University of Barcelona, Spain, August 26th to September 6th, 2019

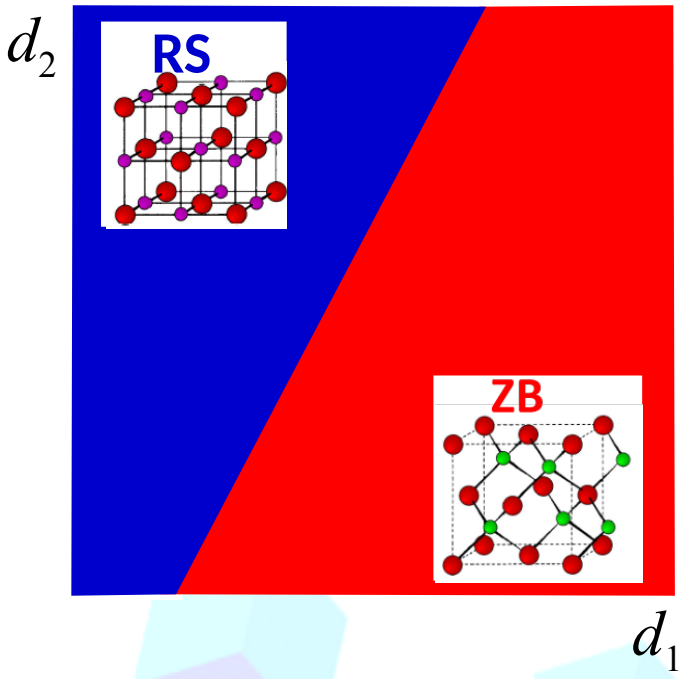
Charts/maps of materials

Charts/maps of materials

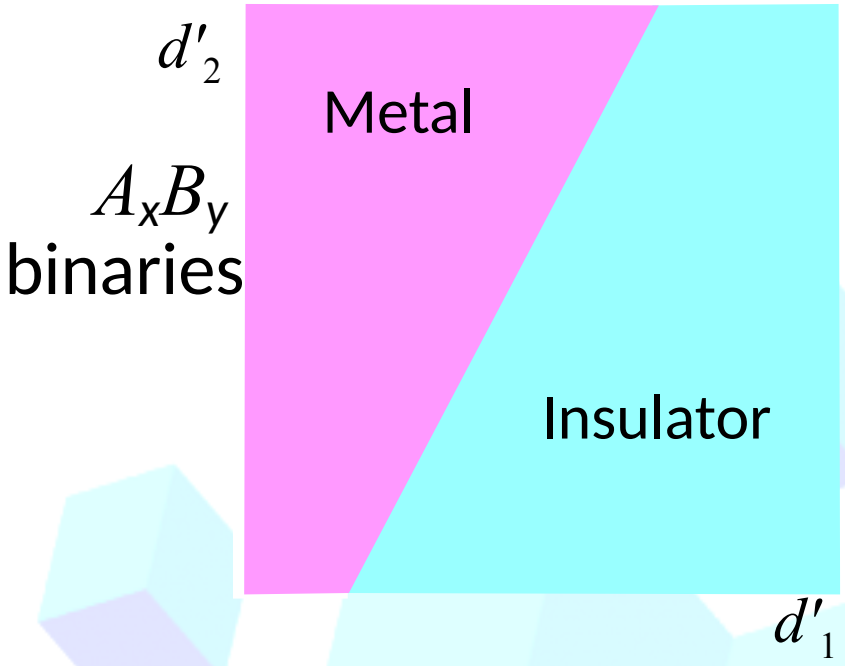


Octet binaries

Charts/maps of materials

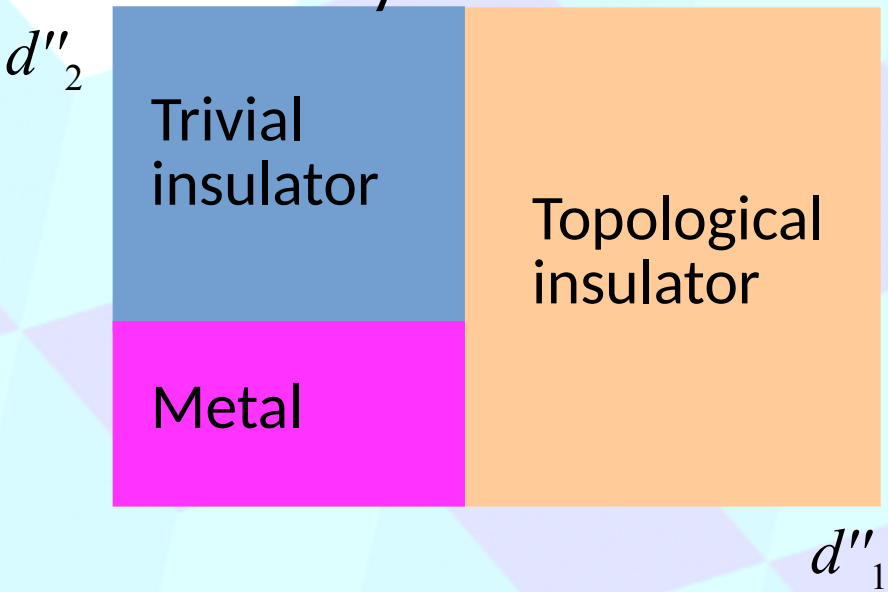


Octet binaries

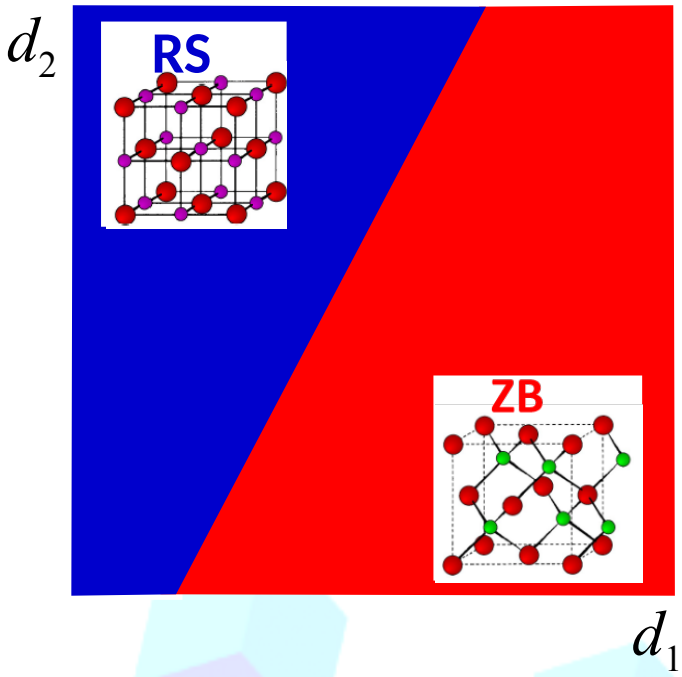


A_xB_y binaries

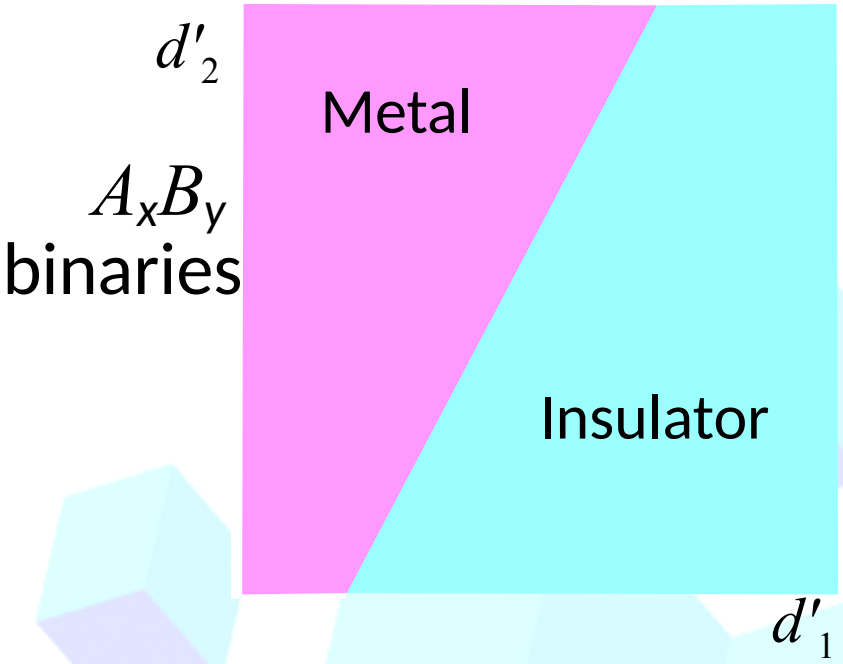
2D honeycomb materials



Charts/maps of materials

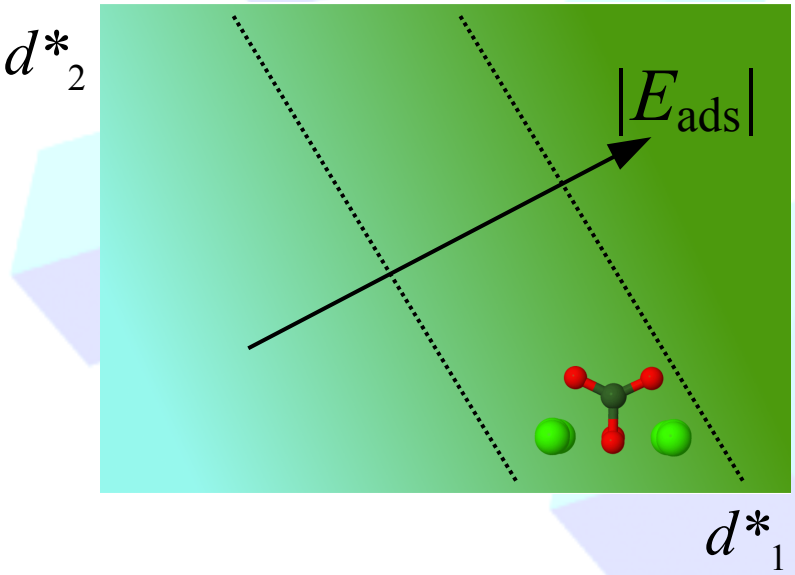


Octet binaries

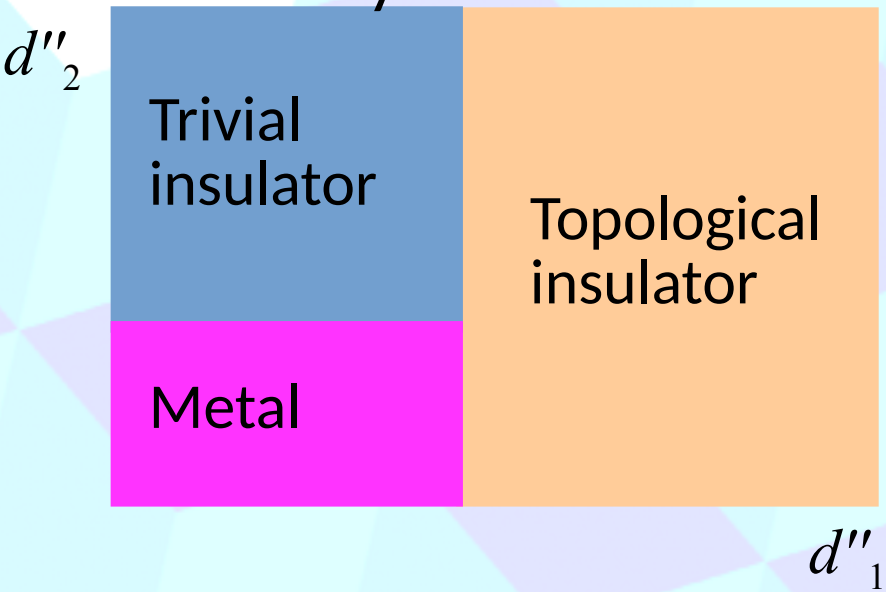


A_xB_y binaries

$E_{ads}(CO_2)$ on oxides



2D honeycomb materials



Building maps of materials properties
A quantum many-body problem

From the Hamiltonian of a physical system, in principle we can derive all properties (observables).

Building maps of materials properties
A quantum many-body problem

From the Hamiltonian of a physical system, in principle we can derive all properties (observables).
But in practice, the Hamiltonian is often not the starting point.

Building maps of materials properties
A quantum many-body problem

From the Hamiltonian of a physical system, in principle we can derive all properties (observables).
But in practice, the Hamiltonian is often not the starting point.

For instance, given a class of chemical compositions
(e.g., via prototype formula, such as ABX_3):

Building maps of materials properties
A quantum many-body problem

From the Hamiltonian of a physical system, in principle we can derive all properties (observables).
But in practice, the Hamiltonian is often not the starting point.

For instance, given a class of chemical compositions (e.g., via prototype formula, such as ABX_3):

- what is the most stable crystal structure of each material in the class?

Building maps of materials properties
A quantum many-body problem

From the Hamiltonian of a physical system, in principle we can derive all properties (observables).
But in practice, the Hamiltonian is often not the starting point.

For instance, given a class of chemical compositions
(e.g., via prototype formula, such as ABX_3):

- what is the most stable crystal structure of each material in the class?
- which materials are metals / topological insulators / superconductors ?
- which material has the highest melting point?
- which materials has a surface optimal for catalysing some chemical reaction?

The Big Picture

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

The Big Picture

- Design of new materials:
  preparation, synthesis, and characterization is complex and costly
- About 240000 *inorganic* materials are known to exist (Springer Materials)

The Big Picture

- Design of new materials:
  preparation, synthesis, and characterization is complex and costly
- About 240000 *inorganic* materials are known to exist (Springer Materials)
- Basic properties determined for very few of them

The Big Picture

- Design of new materials:
  preparation, synthesis, and characterization is complex and costly
- About 240000 *inorganic* materials are known to exist (Springer Materials)
- Basic properties determined for very few of them
- Number of possible materials: practically infinite

The Big Picture

- Design of new materials:
  preparation, synthesis, and characterization is complex and costly
- About 240000 *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

The Big Picture

- Design of new materials:
  preparation, synthesis, and characterization is complex and costly
- About 240000 *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

We have a dream

From the **periodic table of the elements**
to **charts of materials**

We have a dream

From the **periodic table of the elements**
to **charts of materials**

Reihen	Gruppe I. — R^2O	Gruppe II. — RO	Gruppe III. — R^2O^3	Gruppe IV. RH^4 RO^2	Gruppe V. RH^3 R^2O^5	Gruppe VI. RH^2 RO^3	Gruppe VII. RH R^2O^7	Gruppe VIII. — RO^4
1	H=1							
2	Li=7	Be=9.4	B=11	C=12	N=14	O=16	F=19	
3	Na=23	Mg=24	Al=27.3	Si=28	P=31	S=32	Cl=35.5	
4	K=39	Ca=40	—=44	Ti=48	V=51	Cr=52	Mn=55	Fe=56, Co=59, Ni=59, Cu=63.
5	(Cu=63)	Zn=65	—=68	—=72	As=75	Se=78	Br=80	
6	Rb=85	Sr=87	?Yt=88	Zr=90	Nb=94	Mo=96	—=100	Ru=104, Rh=104, Pd=106, Ag=108.
7	(Ag=108)	Cd=112	In=113	Sn=118	Sb=122	Te=125	J=127	
8	Cs=133	Ba=137	?Di=138	?Ce=140	—	—	—	— — — —
9	(—)	—	—	—	—	—	—	
10	—	—	?Er=178	?La=180	Ta=182	W=184	—	Os=195, Ir=197, Pt=198, Au=199.
11	(Au=199)	Hg=200	Tl=204	Pb=207	Bi=208	—	—	
12	—	—	—	Th=231	—	U=240	—	— — — —

Mendeleev's **1871** periodic table

We have a dream

From the **periodic table of the elements**
to **charts of materials**

Reihen	Gruppe I. — R^2O	Gruppe II. — RO	Gruppe III. — R^2O^3	Gruppe IV. RH^4 RO^2	Gruppe V. RH^3 R^2O^5	Gruppe VI. RH^2 RO^3	Gruppe VII. RH R^2O^7	Gruppe VIII. — RO^4
1	H=1							
2	Li=7	Be=9.4	B=11	C=12	N=14	O=16	F=19	
3	Na=23	Mg=24	Al=27.3	Si=28	P=31	S=32	Cl=35.5	
4	K=39	Ca=40	—=44	Ti=48	V=51	Cr=52	Mn=55	Fe=56, Co=59, Ni=59, Cu=63.
5	(Cu=63)	Zn=65	—=68	—=72	As=75	Se=78	Br=80	
6	Rb=85	Sr=87	?Yt=88	Zr=90	Nb=94	Mo=96	—=100	Ru=104, Rh=104, Pd=106, Ag=108.
7	(Ag=108)	Cd=112	In=113	Sn=118	Sb=122	Te=125	J=127	
8	Cs=133	Ba=137	?Di=138	?Ce=140	—	—	—	— — — —
9	(—)	—	—	—	—	—	—	
10	—	—	?Er=178	?La=180	Ta=182	W=184	—	Os=195, Ir=197, Pt=198, Au=199.
11	(Au=199)	Hg=200	Tl=204	Pb=207	Bi=208	—	—	
12	—	—	—	Th=231	—	U=240	—	— — — —

Mendeleev's **1871** periodic table

We have a dream

From the **periodic table of the elements**
to **charts of materials**

Reihen	Gruppe I. — R^2O	Gruppe II. — RO	Gruppe III. — R^2O^3	Gruppe IV. RH^4 RO^2	Gruppe V. RH^3 R^2O^5	Gruppe VI. RH^2 RO^3	Gruppe VII. RH R^2O^7	Gruppe VIII. — RO^4
1	H=1							
2	Li=7	Be=9.4	B=11	C=12	N=14	O=16	F=19	
3	Na=23	Mg=24	Al=27.3	Si=28	P=31	S=32	Cl=35.5	
4	K=39	Ca=40	—=44	Ti=48	V=51	Cr=52	Mn=55	Fe=56, Co=59, Ni=59, Cu=63.
5	(Cu=63)	Zn=65	—=68	—=72	As=75	Se=78	Br=80	
6	Rb=85	Sr=87	?Yt=88	Zr=90	Nb=94	Mo=96	—=100	Ru=104, Rh=104, Pd=106, Ag=108.
7	(Ag=108)	Cd=112	In=113	Sn=118	Sb=122	Te=125	J=127	
8	Cs=133	Ba=137	?Di=138	?Ce=140	—	—	—	— — — —
9	(—)	—	—	—	—	—	—	
10	—	—	?Er=178	?La=180	Ta=182	W=184	—	Os=195, Ir=197, Pt=198, Au=199.
11	(Au=199)	Hg=200	Tl=204	Pb=207	Bi=208	—	—	
12	—	—	—	Th=231	—	U=240	—	— — — —

Ga=69.7 Ge=72.6

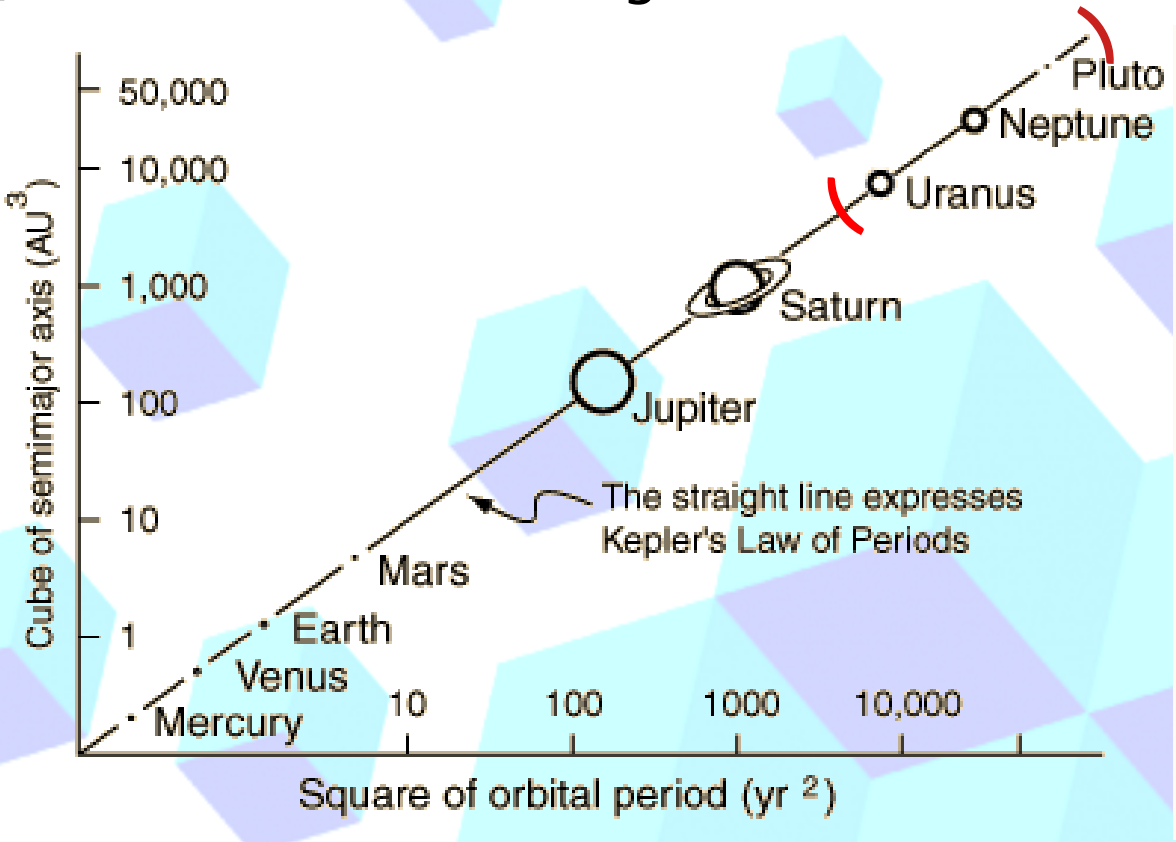
Mendeleev's **1871** periodic table

Learning → Discovery

Suppose
to know the trajectories of all planets in the solar system,
from accurate observations (experiment)
or
by numerically integrating general relativity equations
(calculations at the highest level of theory)

Learning → Discovery

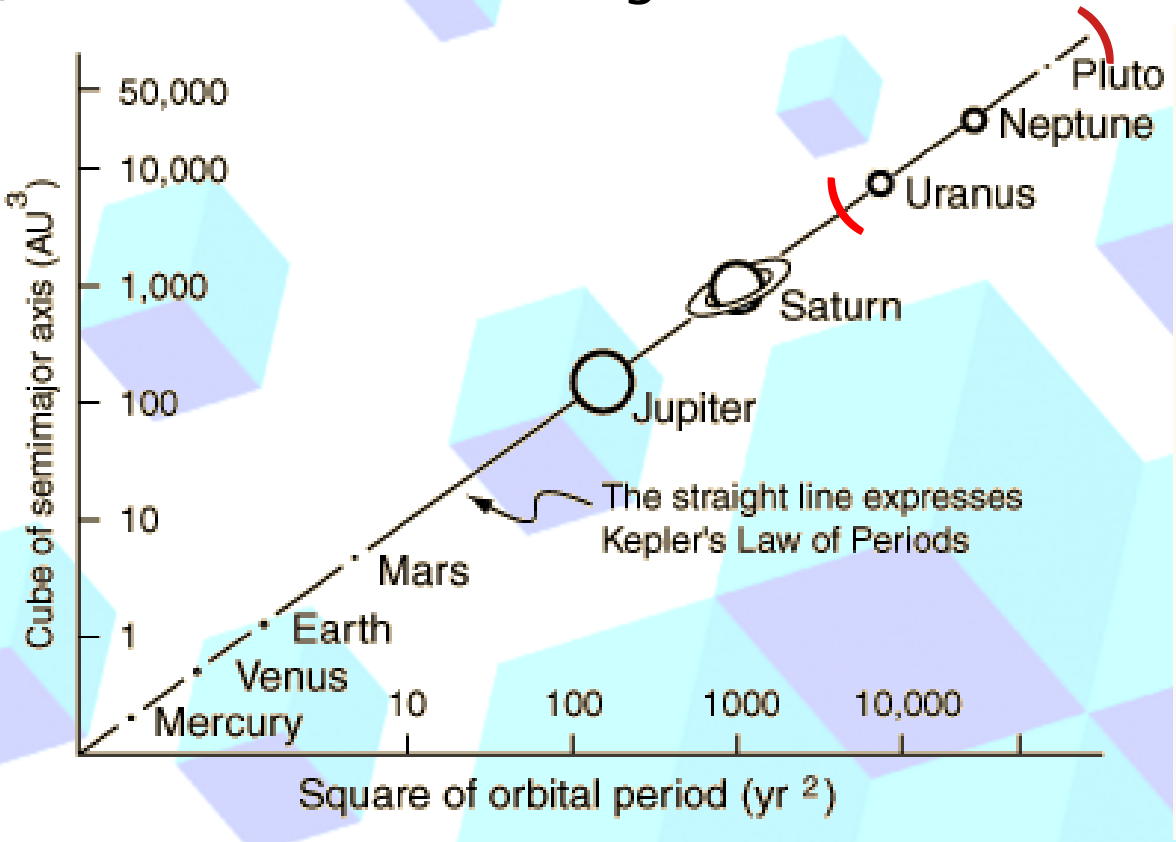
Suppose
to know the trajectories of all planets in the solar system,
from accurate observations (experiment)
or
by numerically integrating general relativity equations
(calculations at the highest level of theory)



(Orbital period)² = C (orbit's major axis)³

Learning → Discovery

Suppose
to know the trajectories of all planets in the solar system,
from accurate observations (experiment)
or
by numerically integrating general relativity equations
(calculations at the highest level of theory)



(Orbital period)² = C (orbit's major axis)³

Data
(collected by
Tycho Brahe)

↓

Statistical learning
(performed by
Johannes Kepler)

↓

Physical law
(assessed by
Isaac Newton)

Supervised (big-)data analysis: a flow chart

Training set

Calculate properties and functions P_i, for many *materials*, i

E.g., Density-Functional Theory

Supervised (big-)data analysis: a 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:
$| i | \boldsymbol{d}_i | P_i |$

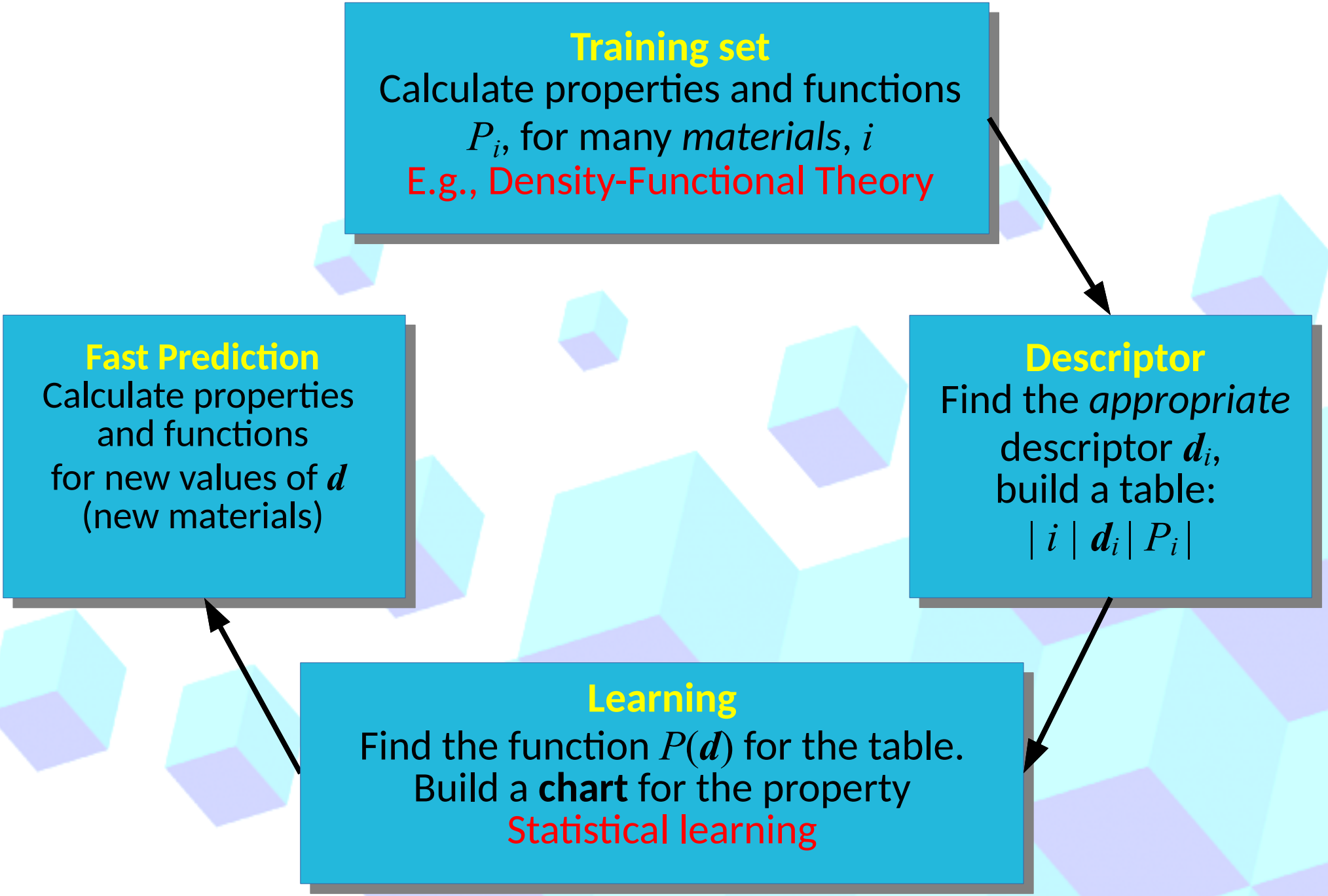
Supervised (big-)data analysis: a flow chart

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

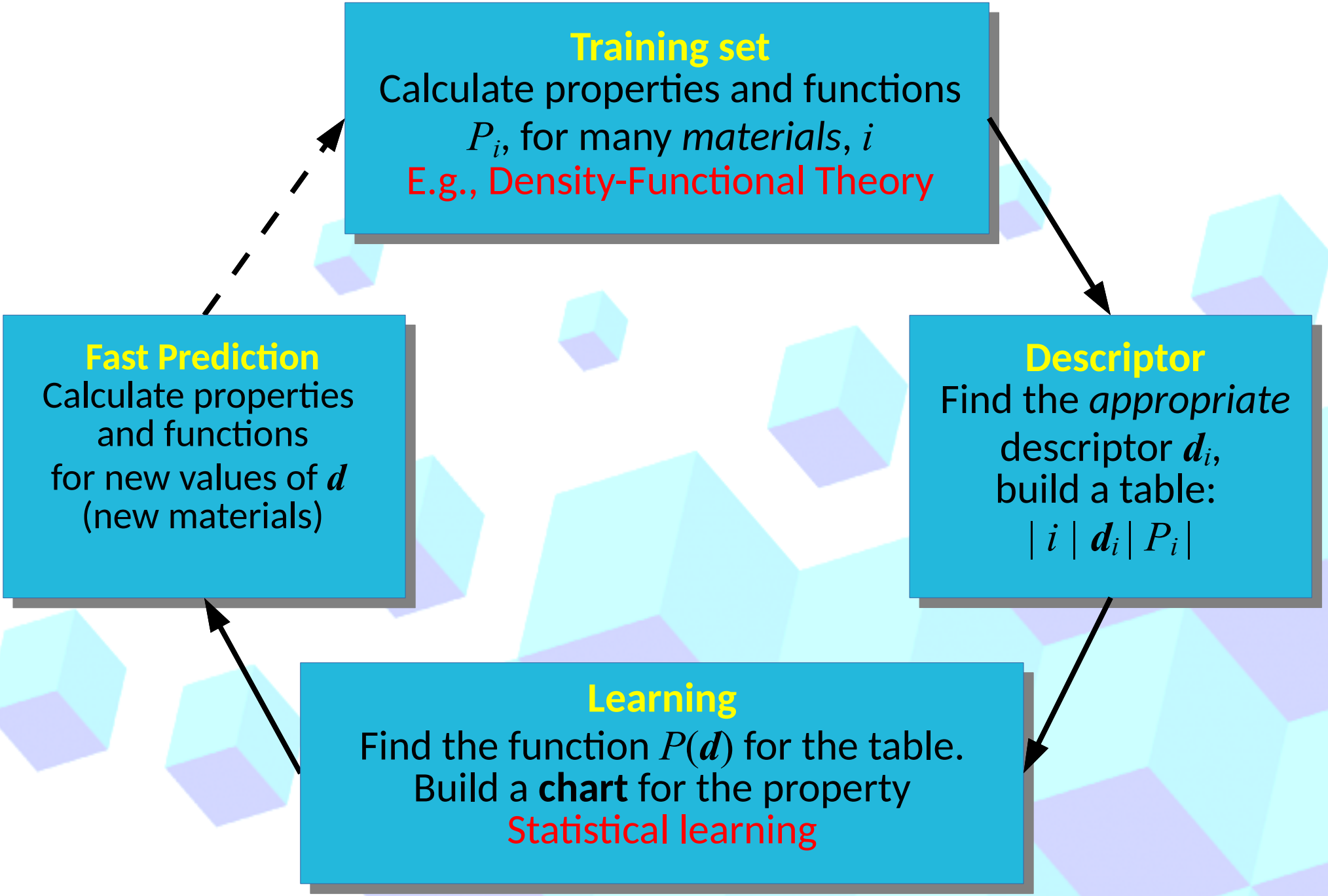
Descriptor
Find the *appropriate*
descriptor $\boldsymbol{d}_i$,
build a table:
$| i | \boldsymbol{d}_i | P_i |$

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

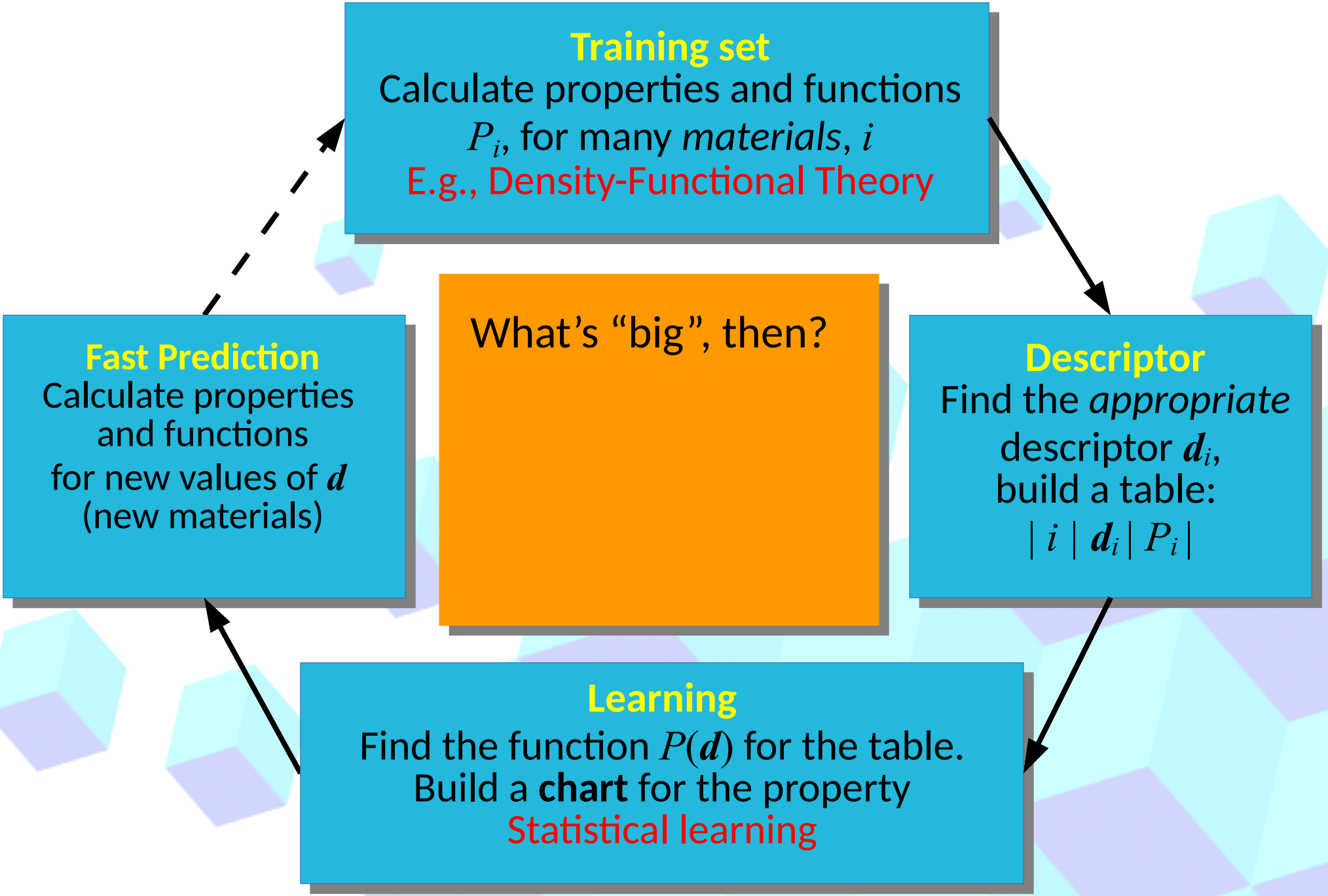
Supervised (big-)data analysis: a flow chart



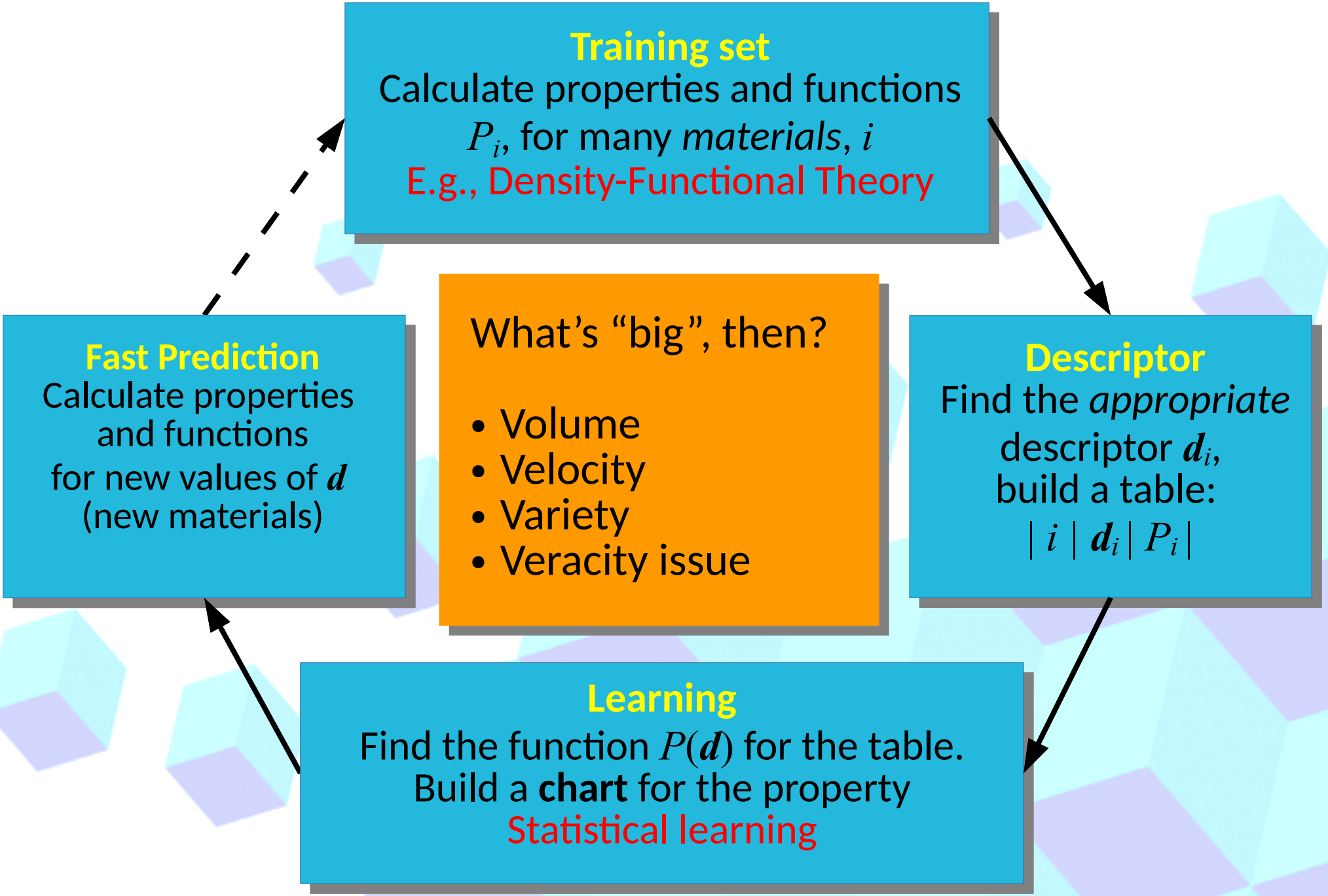
Supervised (big-)data analysis: a flow chart



Supervised big-data analysis: a flow chart



Supervised big-data analysis: a flow chart



Descriptor? Don't we know it from the start?

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

$\{R_\mathrm{I}, Z_\mathrm{I}\} \rightarrow$ Hamiltonian

$\{R_\mathrm{I}\} \rightarrow$ Geometry
- translational, rotational, permutational invariant
- coarse graining $\{R_\mathrm{I}\}$?

$\{Z_\mathrm{I}\} \rightarrow$ Chemistry

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

Regression: Mathematical formulation

Figure of merit to be optimized:

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \sum_{j=1}^{N}\left(P_j - \sum_{l=1}^{M} d_{j,l}c_l\right)^2 = \underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2$$

ℓ_2 norm

Ridge Regression: Mathematical formulation

Figure of merit to be optimized:

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \sum_{j=1}^{N}\left(P_j - \sum_{l=1}^{M} d_{j,l}c_l\right)^2 = \underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2$$

ℓ_2 norm

Regularization (prefer "lower complexity" in the solution)

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

(Linear) ridge regression

Ridge Regression: Mathematical formulation

Figure of merit to be optimized:

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \sum_{j=1}^{N}\left(P_j - \sum_{l=1}^{M} d_{j,l}c_l\right)^2 = \underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2$$

ℓ_2 norm

Regularization (prefer "lower complexity" in the solution)

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

(Linear) ridge regression

Ridge Regression: Mathematical formulation

Figure of merit to be optimized:

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \sum_{j=1}^{N}\left(P_j - \sum_{l=1}^{M} d_{j,l}c_l\right)^2 = \underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2$$

ℓ_2 norm

Regularization (prefer "lower complexity" in the solution)

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

(Linear) ridge regression

Explicit solver:

$$\boldsymbol{c} = \left(\boldsymbol{D}^\top\boldsymbol{D} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{D}^\top\boldsymbol{P}$$

Alternative view, via Hilbert space representation theorem:

$$\boldsymbol{c} = \sum_j \alpha_j \boldsymbol{d}_j$$

Sum over data points!

Kernel Ridge Regression: Mathematical formulation

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P}-\boldsymbol{D}\boldsymbol{c}\|_2^2+\lambda\|\boldsymbol{c}\|_2^2 \quad \Rightarrow \quad \boldsymbol{c}=\left(\boldsymbol{D}^\top\boldsymbol{D}+\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{D}^\top\boldsymbol{P}$$

$$\Downarrow$$

$$\boldsymbol{c}=\sum_j \alpha_j \boldsymbol{d}_j$$

Kernel Ridge Regression: Mathematical formulation

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P}-\boldsymbol{D}\boldsymbol{c}\|_2^2+\lambda\|\boldsymbol{c}\|_2^2 \quad \Rightarrow \quad \boldsymbol{c}=\left(\boldsymbol{D}^\top\boldsymbol{D}+\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{D}^\top\boldsymbol{P}$$

$$\Downarrow$$

$$\boldsymbol{c}=\sum_j \alpha_j \boldsymbol{d}_j$$

$$K_{ij}=\langle \boldsymbol{d}_i,\boldsymbol{d}_j\rangle \quad \textbf{Linear kernel}$$

$$\Downarrow$$

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

Kernel Ridge Regression: Mathematical formulation

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P}-\boldsymbol{D}\boldsymbol{c}\|_2^2+\lambda\|\boldsymbol{c}\|_2^2 \quad \Rightarrow \quad \boldsymbol{c}=\left(\boldsymbol{D}^\top\boldsymbol{D}+\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{D}^\top\boldsymbol{P}$$

$$\Downarrow$$

$$\boldsymbol{c}=\sum_j \alpha_j \boldsymbol{d}_j \qquad K_{ij}=\langle \boldsymbol{d}_i,\boldsymbol{d}_j\rangle \quad \textbf{Linear kernel}$$

$$\Downarrow$$

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P}-\boldsymbol{K}\boldsymbol{\alpha}\|_2^2+\lambda\boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha} \quad \Rightarrow \quad \boldsymbol{\alpha}=(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{P}$$

Kernel Ridge Regression: Mathematical formulation

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P}-\boldsymbol{D}\boldsymbol{c}\|_2^2+\lambda\|\boldsymbol{c}\|_2^2 \quad \Rightarrow \quad \boldsymbol{c}=\left(\boldsymbol{D}^\top\boldsymbol{D}+\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{D}^\top\boldsymbol{P}$$

$$\Downarrow$$

$$\boldsymbol{c}=\sum_j \alpha_j \boldsymbol{d}_j \qquad K_{ij}=\langle \boldsymbol{d}_i,\boldsymbol{d}_j\rangle \quad \textbf{Linear kernel}$$

$$\Downarrow$$

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P}-\boldsymbol{K}\boldsymbol{\alpha}\|_2^2+\lambda\boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha} \quad \Rightarrow \quad \boldsymbol{\alpha}=(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{P}$$

Non-linear kernel

$$\boldsymbol{c}=\sum_j \alpha_j \Phi(\boldsymbol{d}_j) \qquad K_{ij}=\langle \Phi(\boldsymbol{d}_i),\Phi(\boldsymbol{d}_j)\rangle$$

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P}-\boldsymbol{K}\boldsymbol{\alpha}\|_2^2+\lambda\boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha} \quad \Rightarrow \quad \boldsymbol{\alpha}=(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{P}$$

Kernel Ridge Regression: Mathematical formulation

Non-linear kernel

$$\boldsymbol{c} = \sum_j \alpha_j \Phi(\boldsymbol{d}_j) \qquad K_{ij} = \langle \Phi(\boldsymbol{d}_i), \Phi(\boldsymbol{d}_j) \rangle$$

$$\underset{\boldsymbol{c} \in \mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P} - \boldsymbol{K}\boldsymbol{\alpha}\|_2^2 + \lambda \boldsymbol{\alpha}^\top \boldsymbol{K} \boldsymbol{\alpha} \quad \Rightarrow \quad \boldsymbol{\alpha} = (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{P}$$

$K_{ij} = \langle \boldsymbol{d}_i, \boldsymbol{d}_j \rangle$ **Linear kernel**

$K_{ij} = (\langle \boldsymbol{d}_i, \boldsymbol{d}_j \rangle + b)^n$ **Polynomial kernel**

$K_{ij} = \exp\left(\frac{\|\boldsymbol{d}_i - \boldsymbol{d}_j\|^2}{2\sigma^2}\right)$ **Gaussian (radial basis function) kernel**

$K_{ij} = \exp\left(\frac{\|\boldsymbol{d}_i - \boldsymbol{d}_j\|}{\sigma}\right)$ **Laplacian kernel**

Kernel Ridge Regression: Mathematical formulation

Non-linear kernel

$$\boldsymbol{c} = \sum_j \alpha_j \Phi(\boldsymbol{d}_j) \qquad K_{ij} = \langle \Phi(\boldsymbol{d}_i), \Phi(\boldsymbol{d}_j) \rangle$$

$$\underset{\boldsymbol{c} \in \mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P} - \boldsymbol{K}\boldsymbol{\alpha}\|_2^2 + \lambda \boldsymbol{\alpha}^\top \boldsymbol{K} \boldsymbol{\alpha} \quad \Rightarrow \quad \boldsymbol{\alpha} = (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{P}$$

$$K_{ij} = \langle \boldsymbol{d}_i, \boldsymbol{d}_j \rangle$$

$$K_{ij} = (\langle \boldsymbol{d}_i, \boldsymbol{d}_j \rangle + b)^n$$

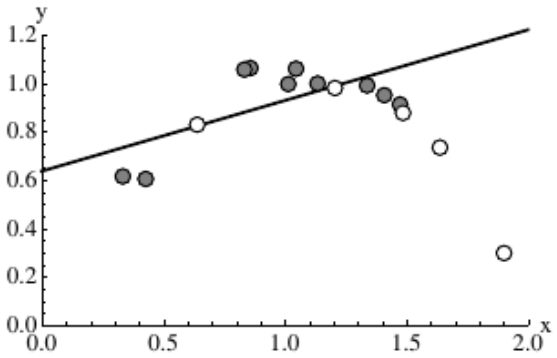
$$K_{ij} = \exp\left(\frac{\|\boldsymbol{d}_i - \boldsymbol{d}_j\|^2}{2\sigma^2}\right)$$

$$K_{ij} = \exp\left(\frac{\|\boldsymbol{d}_i - \boldsymbol{d}_j\|}{\sigma}\right)$$

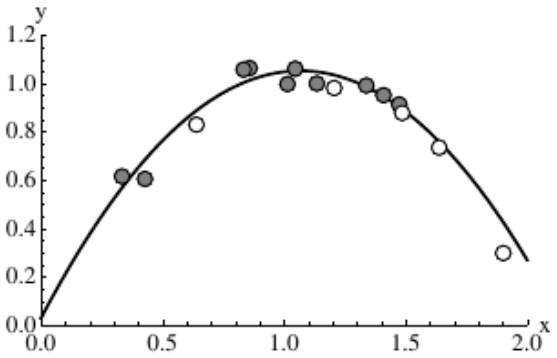
In all cases,
a kernel introduces a
similarity measure

Regularized regression in practice: beware of overfitting

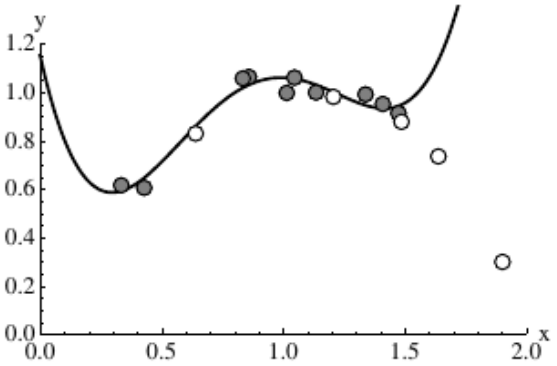
Underfitting



Fitting



Overfitting



Training/ validation error

0.123 / 0.443

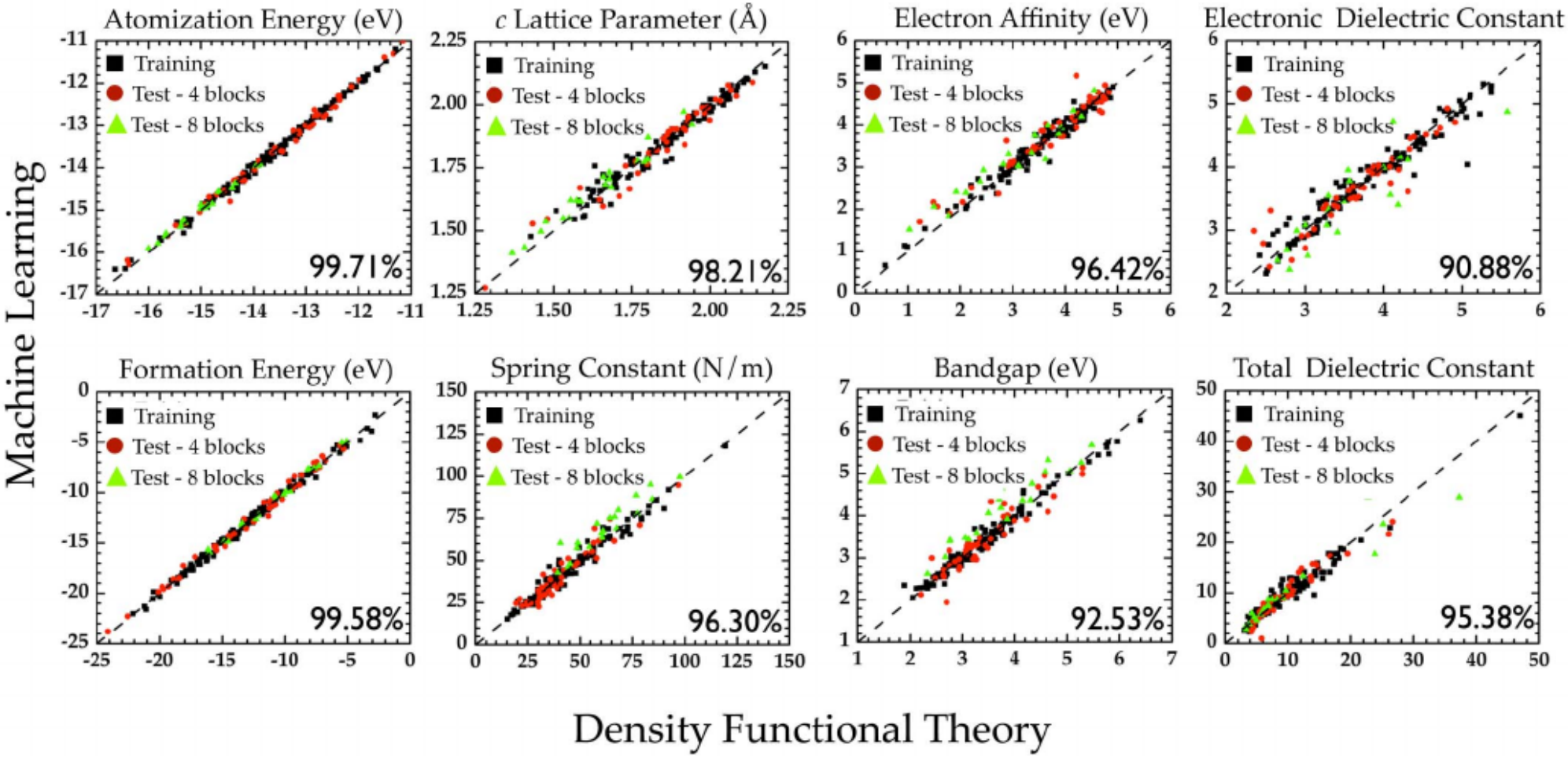
0.044 / 0.068

0.036 / 0.939

KRR success stories: 1D polymers “eugenetics”

Data: 175 linear 4-blocks periodic polymers. 7 blocks: CH_2, SiF_2, $SiCl_2$, GeF_2, $GeCl_2$, SnF_2, $SnCl_2$,

Descriptor: 20 dimensions [# building blocks of type *i*, of *ii* pairs, of *iii* triplets]



Pilania, Wang, …, and Ramprasad, Scientific Reports 3, 2810 (2013). DOI: 10.1038/srep02810

KRR success stories: *n*-grams for kaggle

Research Prediction Competition

Nomad2018 Predicting Transparent Conductors

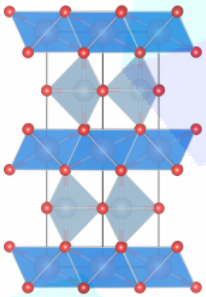
Predict the key properties of novel transparent semiconductors

€5,000
Prize Money

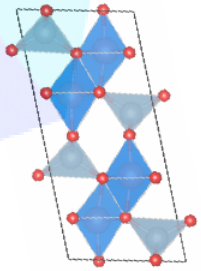
883 teams · 12 days ago

Overview | Data | Kernels | Discussion | Leaderboard | Rules | Team | Host | My Submissions | **Late Submission**

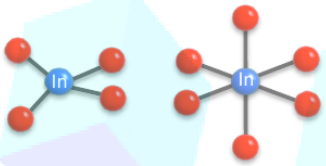
p63/mmc



C2/m

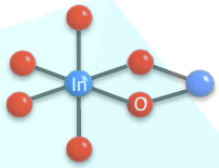


Uni-gram



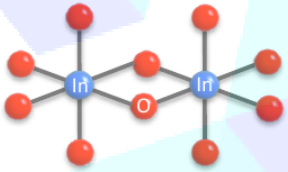
In4 In6

Bi-gram



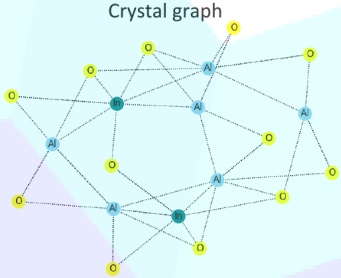
In6 – O2

Tri-gram



In6 – O2 – In6

Crystal graph



Nodes = Atoms

Edges = Bonds

Input features: count number of sequences of various lengths

Unigrams: 2 O3, 2 Ga4, 1 O2, 1 In5

Bigrams: 2 O3-Ga4, 2 Ga4-O2, 1 O3-In5

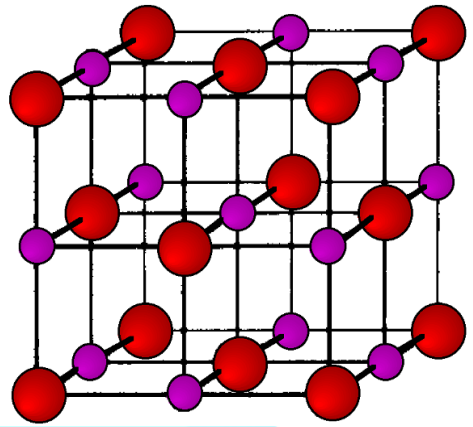
Sutton *et al.*, npj Comp. Materials, in press (2019), arXiv: 1812.00085

An example: predicting crystal structures from the composition

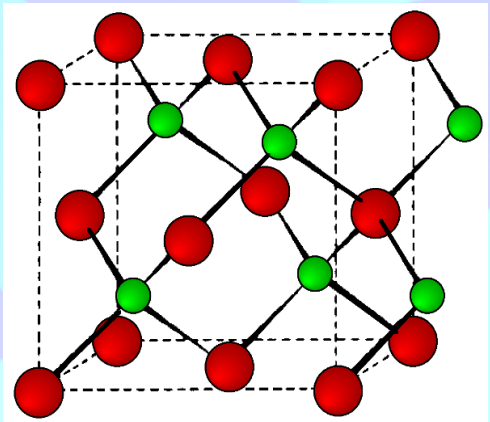
82 octet AB binary compounds

hydrogen 1 H																	helium 2 He
lithium 3 Li	beryllium 4 Be											boron 5 B	carbon 6 C	nitrogen 7 N	oxygen 8 O	fluorine 9 F	neon 10 Ne
sodium 11 Na	magnesium 12 Mg											aluminium 13 Al	silicon 14 Si	phosphorus 15 P	sulfur 16 S	chlorine 17 Cl	argon 18 Ar
potassium 19 K	calcium 20 Ca	scandium 21 Sc	titanium 22 Ti	vanadium 23 V	chromium 24 Cr	manganese 25 Mn	iron 26 Fe	cobalt 27 Co	nickel 28 Ni			gallium 31 Ga	germanium 32 Ge	arsenic 33 As	selenium 34 Se	bromine 35 Br	krypton 36 Kr
rubidium 37 Rb	strontium 38 Sr	yttrium 39 Y	zirconium 40 Zr	niobium 41 Nb	molybdenum 42 Mo	technetium 43 Tc	ruthenium 44 Ru	rhodium 45 Rh	palladium 46 Pd			indium 49 In	tin 50 Sn	antimony 51 Sb	tellurium 52 Te	iodine 53 I	xenon 54 Xe
caesium 55 Cs	barium 56 Ba	57-70 *	lutetium 71 Lu	hafnium 72 Hf	tantalum 73 Ta	tungsten 74 W	rhenium 75 Re	osmium 76 Os	iridium 77 Ir	platinum 78 Pt	gold 79 Au	mercury 80 Hg	thallium 81 Tl	lead 82 Pb	bismuth 83 Bi	polonium 84 Po	astatine 85 At	radon 86 Rn
francium	radium	89-102 **	lawrencium 103 Lr	rutherfordium 104 Rf	dubnium 105 Db	seaborgium 106 Sg	bohrium 107 Bh	hassium 108 Hs	meitnerium 109 Mt	ununnilium 110 Uun	ununuunium 111 Uuu	ununbium 112 Uub		ununquadium 114 Uuq				
copper 29 Cu	zinc 30 Zn																
silver 47 Ag	cadmium 48 Cd																

* Lanthanide series	lanthanum 57 La	cerium 58 Ce	praseodymium 59 Pr	neodymium 60 Nd	promethium 61 Pm	samarium 62 Sm	europium 63 Eu	gadolinium 64 Gd	terbium 65 Tb	dysprosium 66 Dy	holmium 67 Ho	erbium 68 Er	thulium 69 Tm	ytterbium 70 Yb
** Actinide series	actinium 89 Ac	thorium 90 Th	protactinium 91 Pa	uranium 92 U	neptunium 93 Np	plutonium 94 Pu	americium 95 Am	curium 96 Cm	berkelium 97 Bk	californium 98 Cf	einsteinium 99 Es	fermium 100 Fm	mendelevium 101 Md	nobelium 102 No



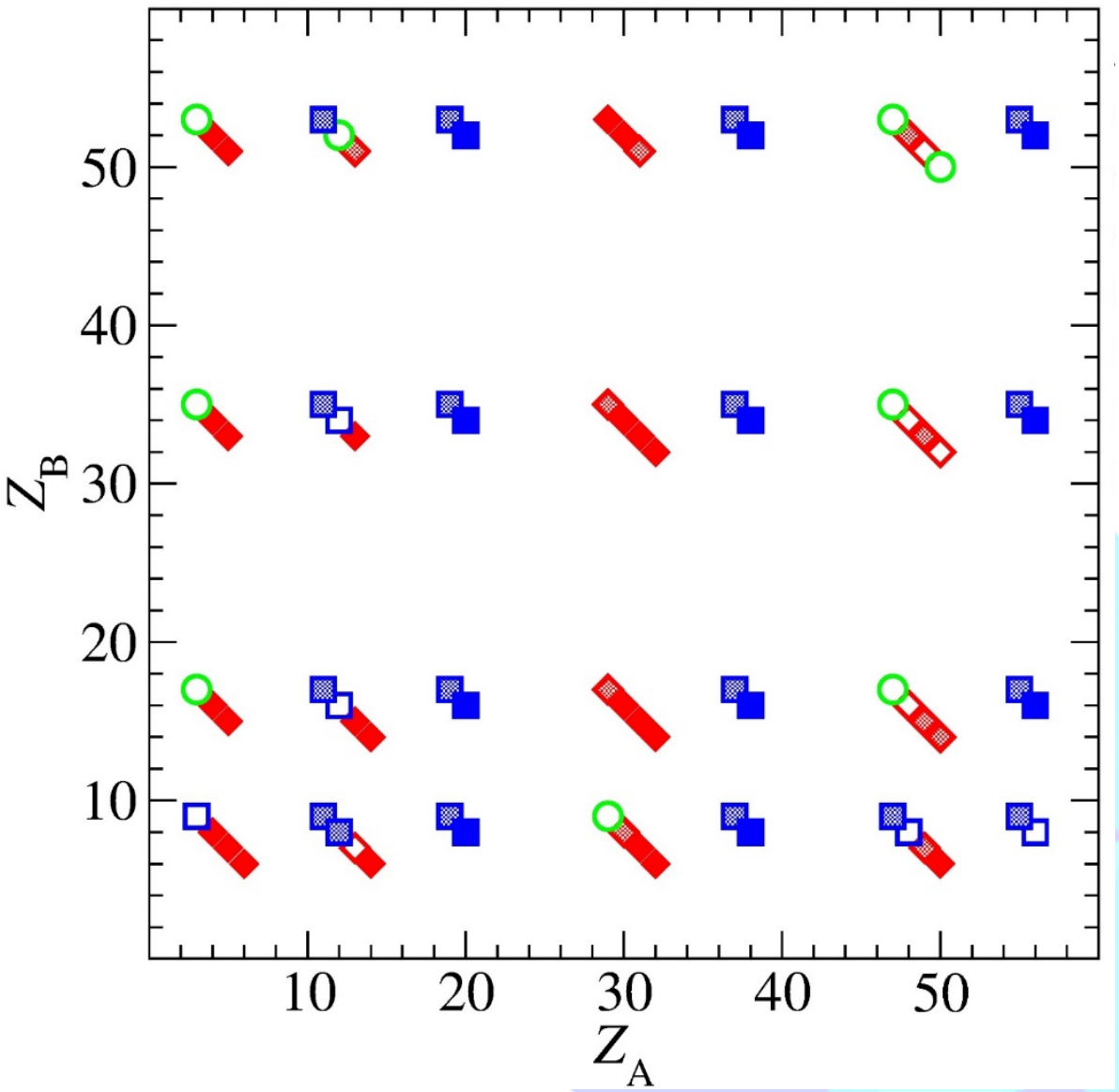
Rock salt



Zinc blende

An example: predicting crystal structures from the composition

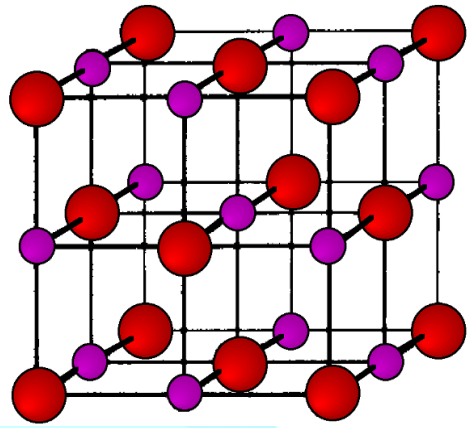
82 octet AB binary compounds



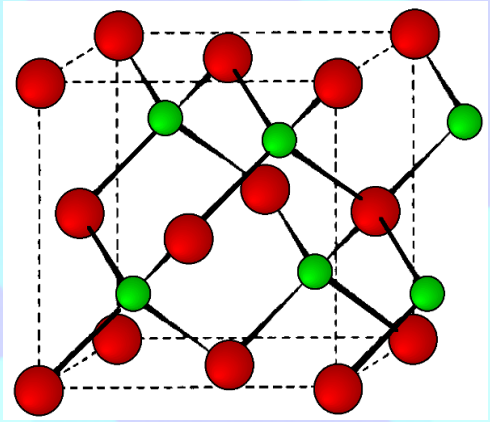
Rock salt

Rock salt/Zinc blende

Zinc blende



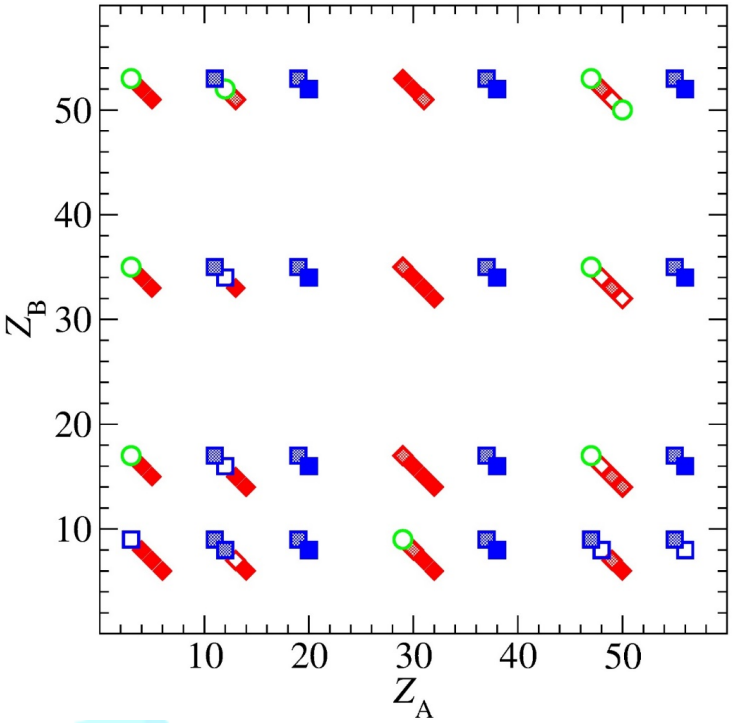
Rock salt



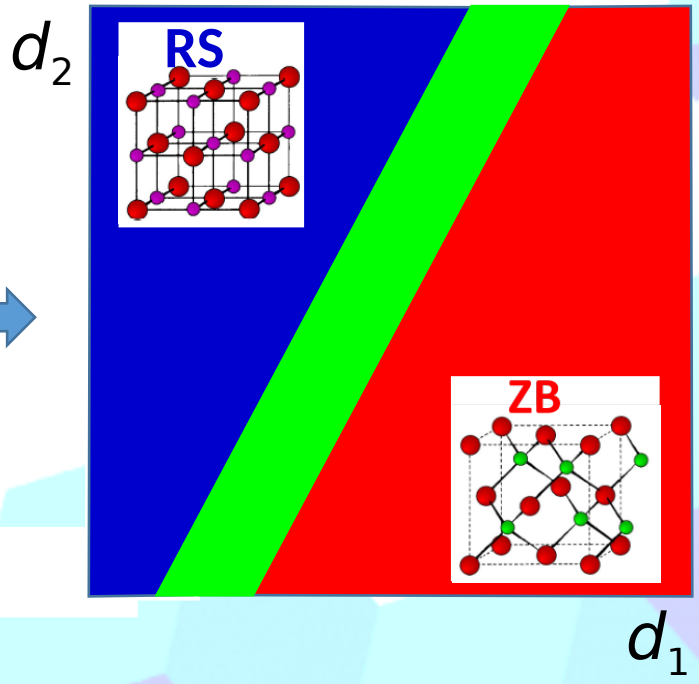
Zinc blende

An example: predicting crystal structures from the composition

82 octet AB binary compounds



?



- Rock salt
- Rock salt/Zinc blende
- Zinc blende

J. A. van Vechten, Phys. Rev. 182, 891 (1969).

J. C. Phillips, Rev. Mod. Phys. 42, 317 (1970).

J. John and A.N. Bloch, Phys. Rev. Lett. 33, 1095 (1974)

J. R. Chelikowsky and J. C. Phillips, Phys. Rev. B 33, 2453 (1978)

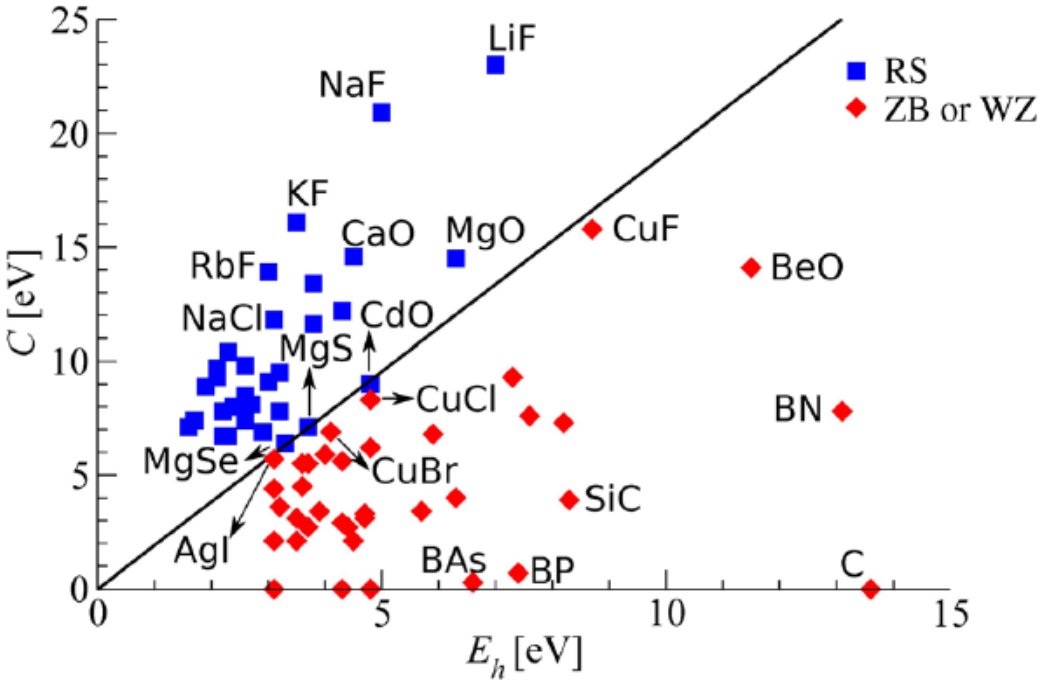
A. Zunger, Phys. Rev. B 22, 5839 (1980).

D. G. Pettifor, Solid State Commun. 51, 31 (1984).

Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, Phys. Rev. B 85, 104104 (2012).

An example: predicting crystal structures from the composition

82 octet AB binary compounds



The descriptor proposed by Phillips and van Vechten in 1969-70 depends on:
- lattice parameter
- electrical conductivity

d_1

J. A. van Vechten, Phys. Rev. 182, 891 (1969).

J. C. Phillips, Rev. Mod. Phys. 42, 317 (1970).

J. John and A.N. Bloch, Phys. Rev. Lett. 33, 1095 (1974).

J. R. Chelikowsky and J. C. Phillips, Phys. Rev. B 33, 2453 (1978).

A. Zunger, Phys. Rev. B 22, 5839 (1980).

D. G. Pettifor, Solid State Commun. 51, 31 (1984).

Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, Phys. Rev. B 85, 104104 (2012).

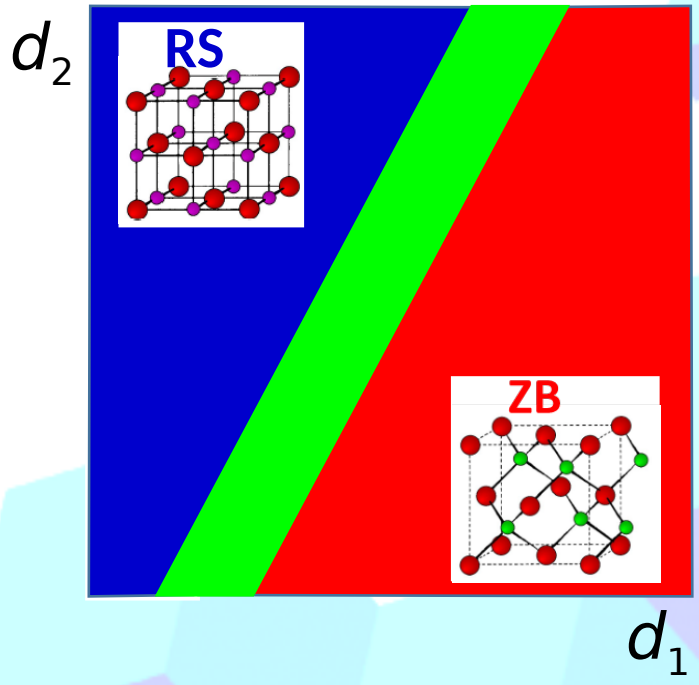
An example: predicting crystal structures from the composition

82 octet AB binary compounds

Ansatz: atomic features

- HOMO
- LUMO
- Ionization Potential
- Electron Affinity
- Radius of valence *s* orbital
- Radius of valence *p* orbital
- Radius of valence *d* orbital
- … ?

E(Rock salt) – *E*(Zinc blende)



Rock salt

Rock salt/Zinc blende

Zinc blende

J. A. van Vechten, Phys. Rev. 182, 891 (1969).

J. C. Phillips, Rev. Mod. Phys. 42, 317 (1970).

J. John and A.N. Bloch, Phys. Rev. Lett. 33, 1095 (1974)

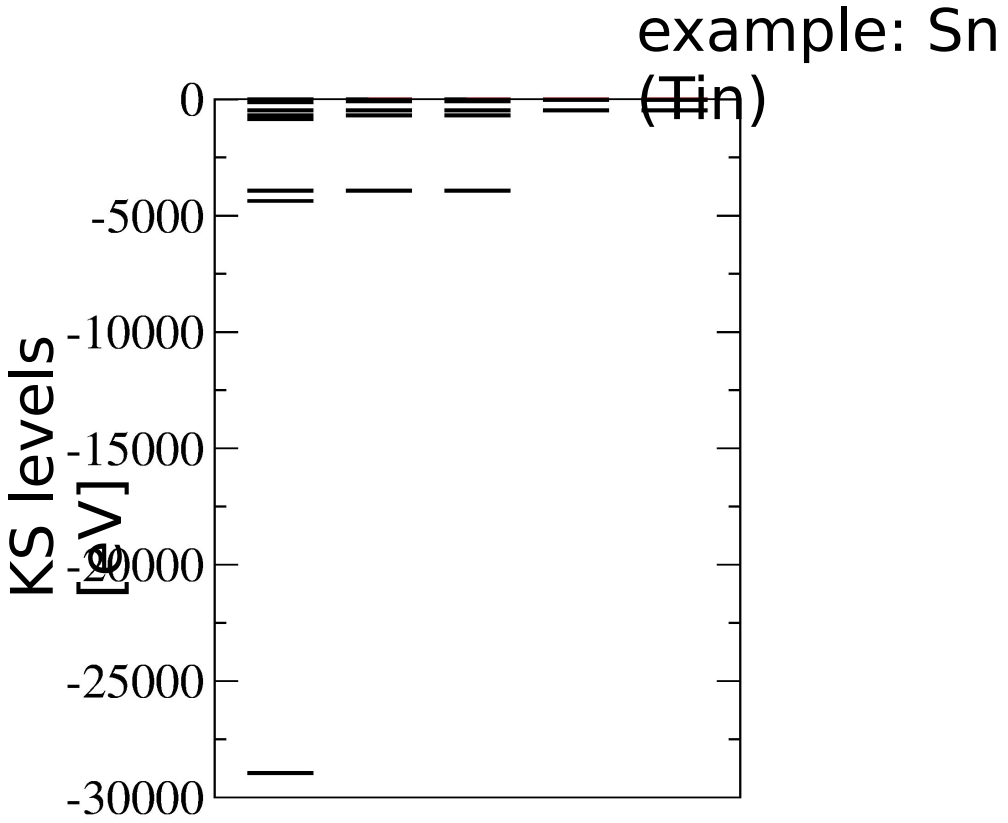
J. R. Chelikowsky and J. C. Phillips, Phys. Rev. B 33, 2453 (1978)

A. Zunger, Phys. Rev. B 22, 5839 (1980).

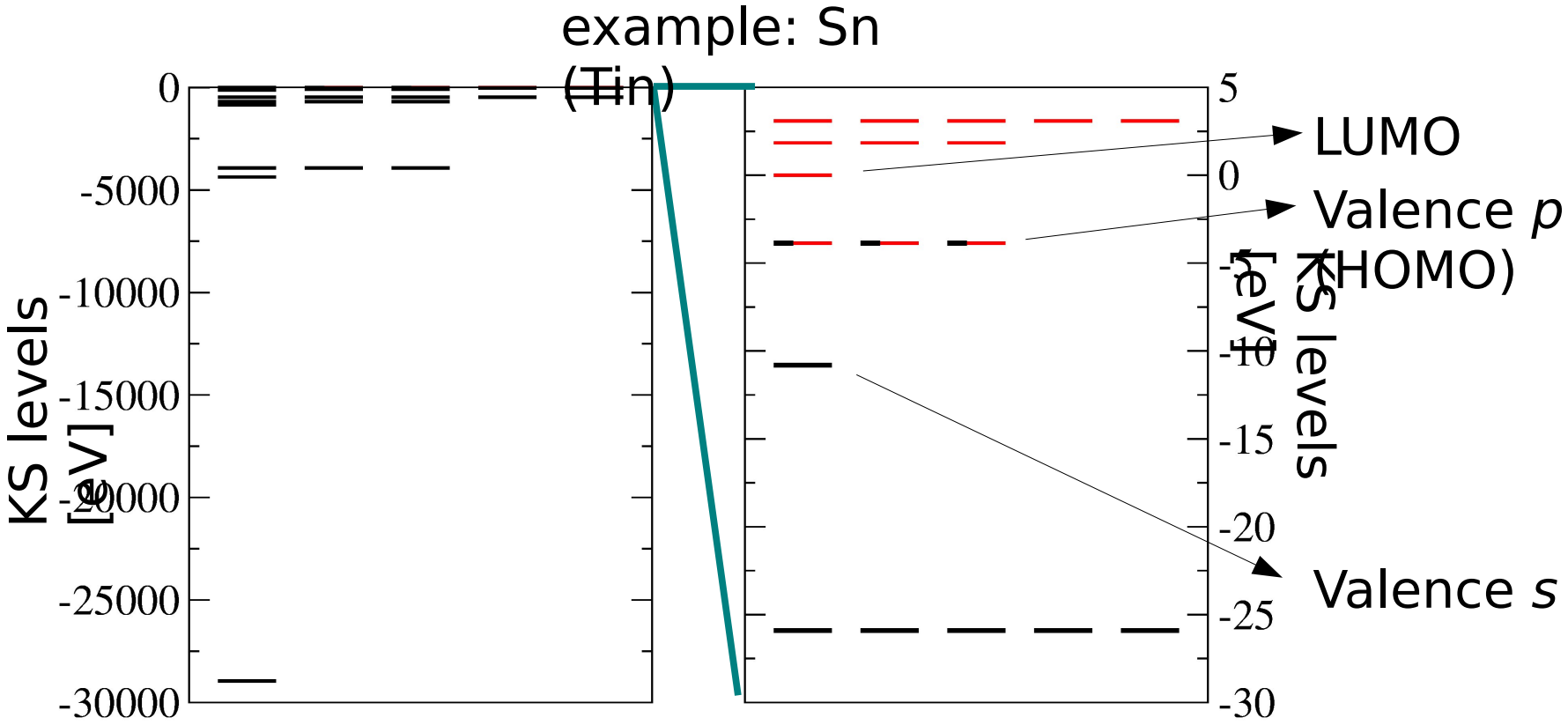
D. G. Pettifor, Solid State Commun. 51, 31 (1984).

Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, Phys. Rev. B 85, 104104 (2012).

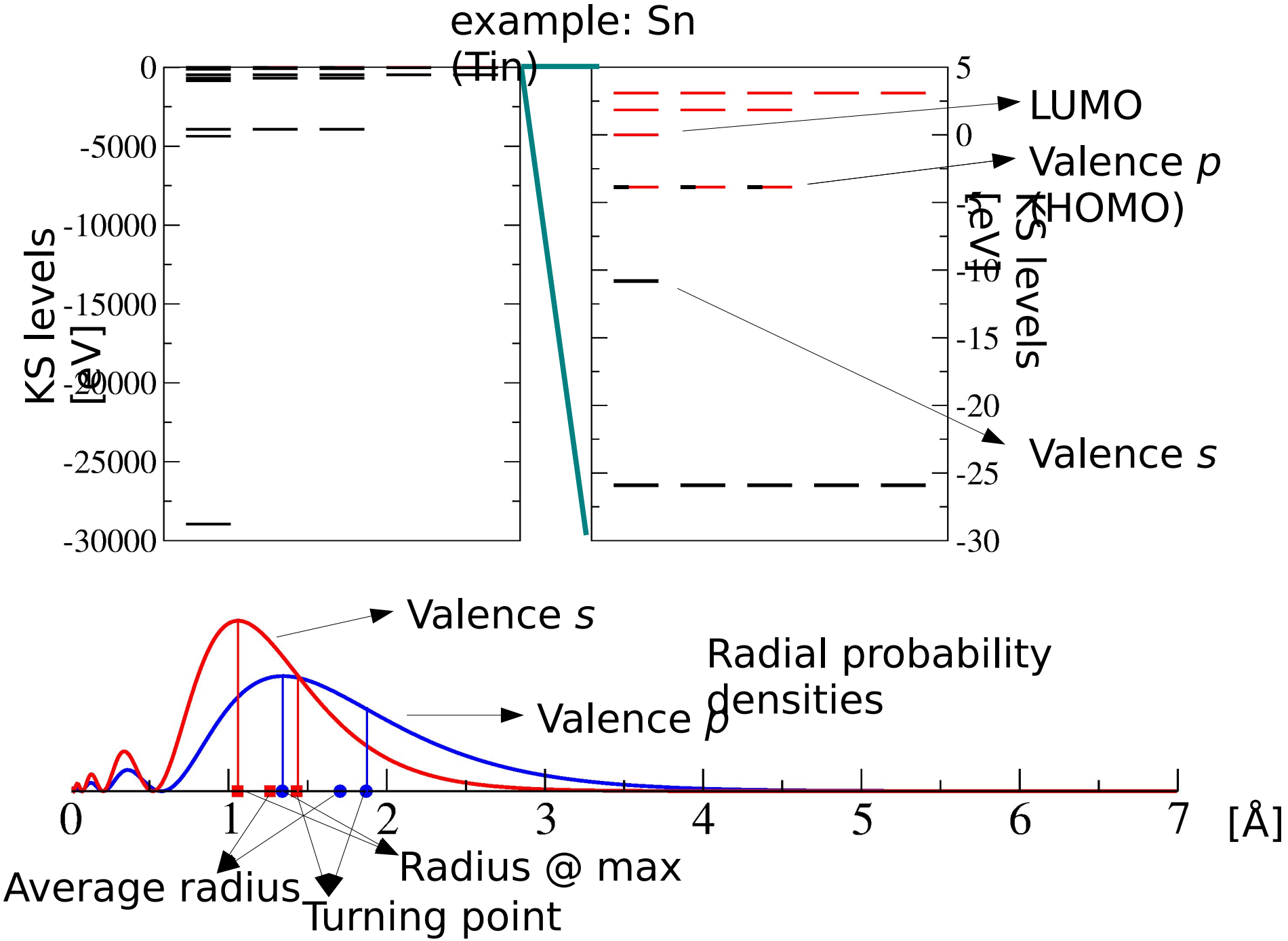
Primary (atomic) features



Primary (atomic) features



Primary (atomic) features



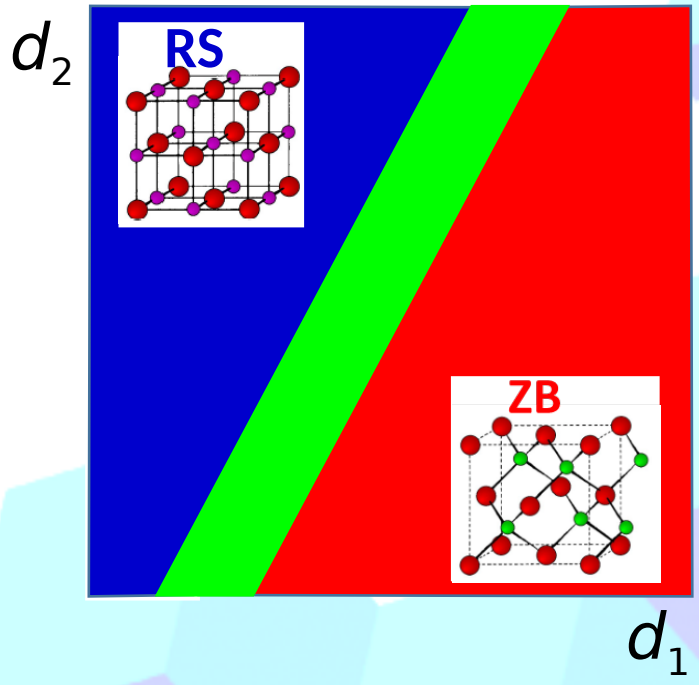
An example: predicting crystal structures from the composition

82 octet AB binary compounds

Ansatz: atomic features

- HOMO
- LUMO
- Ionization Potential
- Electron Affinity
- Radius of valence *s* orbital
- Radius of valence *p* orbital
- Radius of valence *d* orbital
- … ?

E(Rock salt) – *E*(Zinc blende)



Rock salt

Rock salt/Zinc blende

Zinc blende

J. A. van Vechten, Phys. Rev. 182, 891 (1969).

J. C. Phillips, Rev. Mod. Phys. 42, 317 (1970).

J. John and A.N. Bloch, Phys. Rev. Lett. 33, 1095 (1974)

J. R. Chelikowsky and J. C. Phillips, Phys. Rev. B 33, 2453 (1978)

A. Zunger, Phys. Rev. B 22, 5839 (1980).

D. G. Pettifor, Solid State Commun. 51, 31 (1984).

Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, Phys. Rev. B 85, 104104 (2012).

Compressed sensing

Aim: finding descriptors and learning predictive models
Ansatz:
$\boldsymbol{P} = c_1\boldsymbol{d}_1 + c_2\boldsymbol{d}_2 + \ldots c_n\boldsymbol{d}_n$
Where
$\boldsymbol{P}$ is the property of interest
$\boldsymbol{d}_1, \ldots \boldsymbol{d}_n$ are candidate features, i.e., nonlinear functions of primary features (EA, IP, …)
$c_1, \ldots c_n$ are unknown coefficients, with the extra constraint that these (nonzero) coefficients should be as few as possible.

Compressed sensing

With a foreword on
dimensionality reduction

Aim: finding descriptors and learning pre models
Ansatz:
$\boldsymbol{P} = c_1\boldsymbol{d}_1 + c_2\boldsymbol{d}_2 + \ldots c_n\boldsymbol{d}_n$
Where
$\boldsymbol{P}$ is the property of interest
$\boldsymbol{d}_1, \ldots \boldsymbol{d}_n$ are candidate features, i.e., nonlinear functions of primary features (EA, IP, …)
$c_1, \ldots c_n$ are unknown coefficients, with the extra constraint that these (nonzero) coefficients should be as few as possible.

Linear dimensionality reduction: Principal components

Pearson, K. "On Lines and Planes of Closest Fit to Systems of Points in Space". Philosophical Magazine 2, 559 (1901)

Linear dimensionality reduction: Principal components

Pearson, K. "On Lines and Planes of Closest Fit to Systems of Points in Space". Philosophical Magazine 2, 559 (1901)

Orthonormal transformation of coordinates, converting a set of (possibly) linearly correlated coordinates into a new set of linearly uncorrelated (called principal or normal) components, such that the first component has the largest variance and each subsequent has the largest variance constrained to being orthogonal to all the preceding components

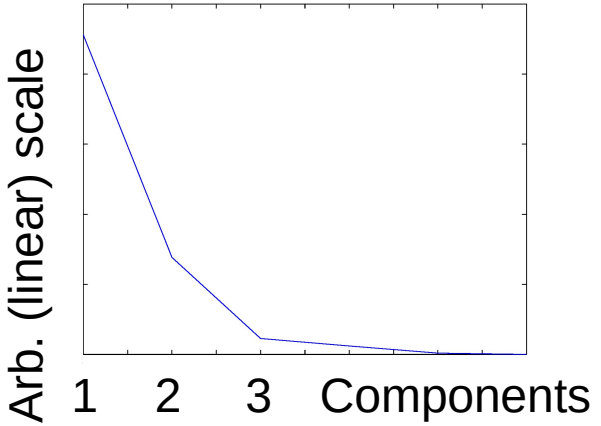
Linear dimensionality reduction: Principal components

Ansatz: atomic features

- Valence number Z_v
- Energy of valence *s* orbital E_s
- Energy of valence *p* orbital E_p
- Radius of valence *s* orbital r_s
- Radius of valence *p* orbital r_p

r_s, r_p, $E_s/\sqrt{Z_v}$, $E_p/\sqrt{z_v}$,
for A and B atoms

normal) components, such that the first component has the largest variance and each subsequent has the largest variance constrained to being orthogonal to all the preceding components



Saad, …, Chelikowsky, and Andreoni, PRB 85, 104104 (2012)

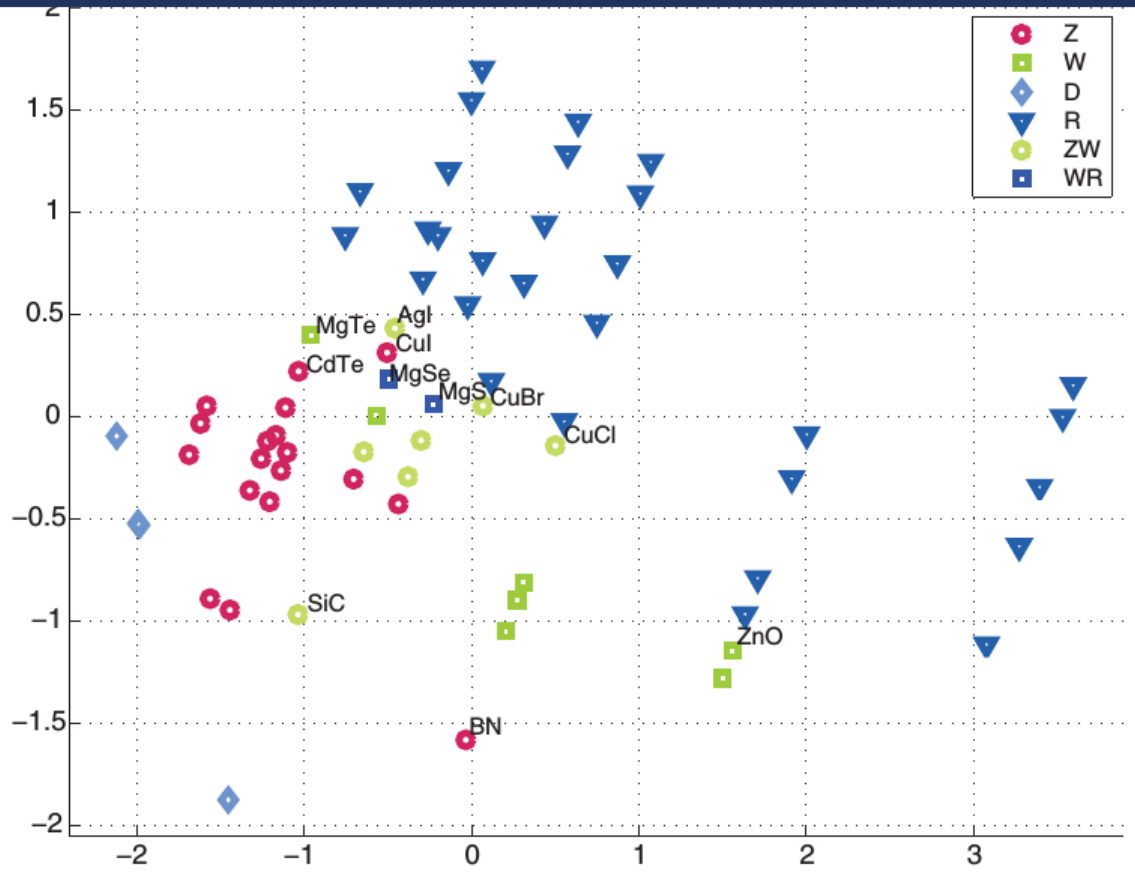
Linear dimensionality reduction: Principal components

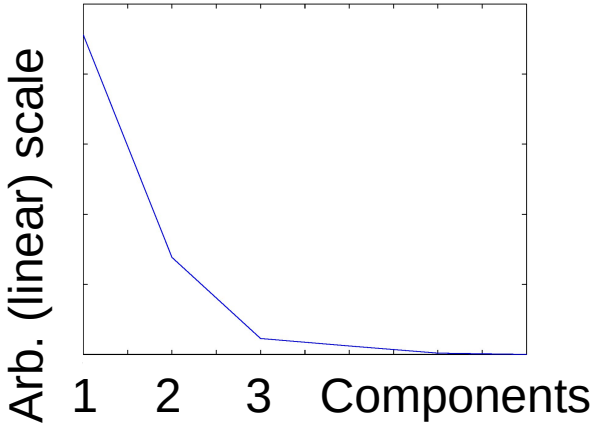
Ansatz: atomic features

- Valence number Z_v
- Energy of valence *s* orbital E_s
- Energy of valence *p* orbital E_p
- Radius of valence *s* orbital r_s
- Radius of valence *p* orbital r_p

r_s, r_p, $E_s/\sqrt{Z_v}$, $E_p/\sqrt{z_v}$,
for A and B atoms

linearly uncorrelated (called principal or normal) components, such that the first component has the largest variance and each subsequent has the largest variance constrained to being orthogonal to all the preceding components





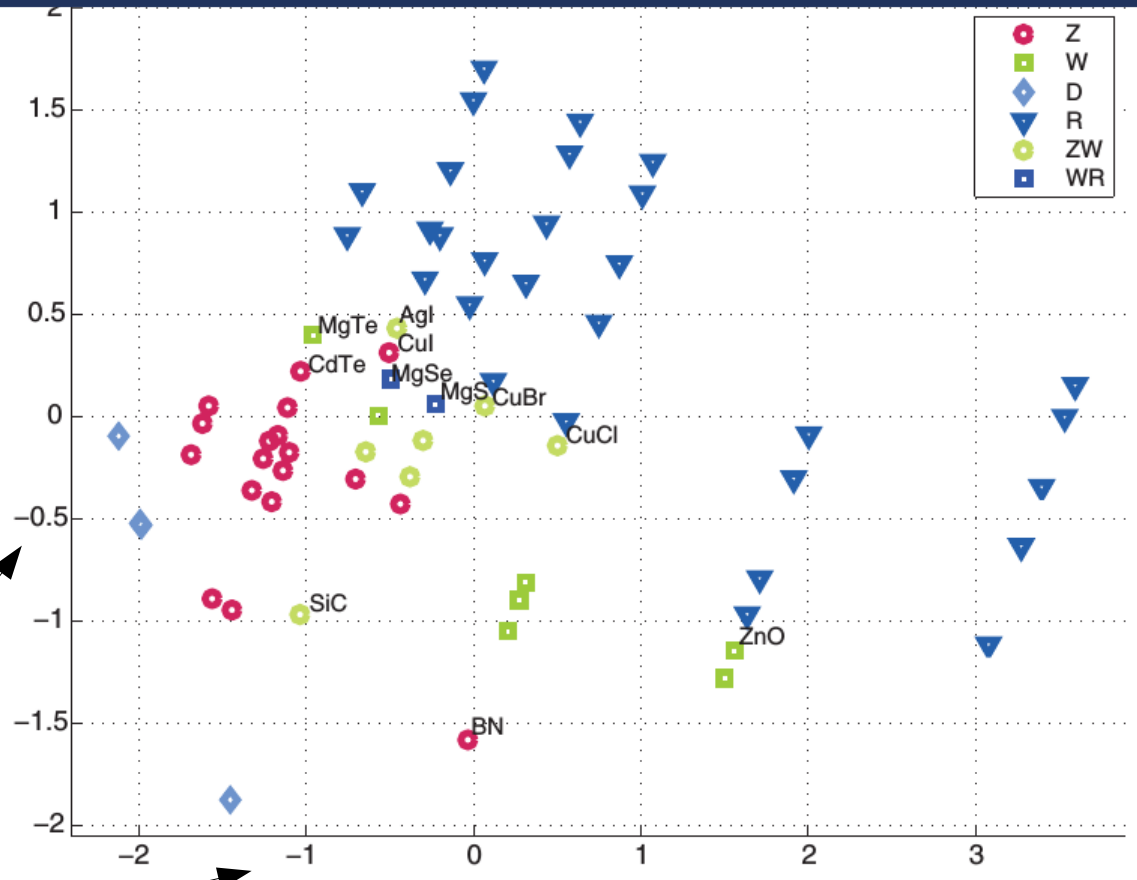
Saad, …, Chelikowsky, and Andreoni, PRB 85, 104104 (2012)

Linear dimensionality reduction: Principal components

Ansatz: atomic features

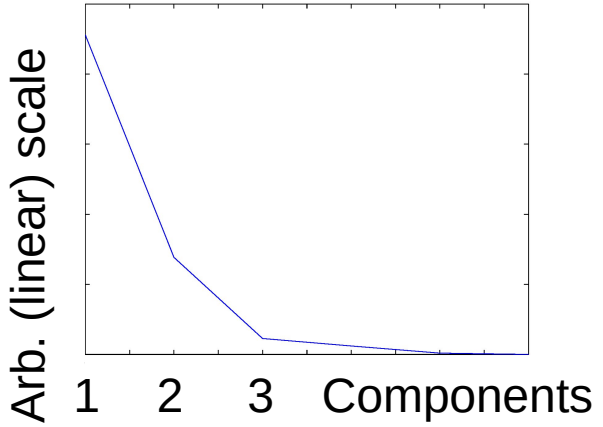
- Valence number Z_v
- Energy of valence *s* orbital E_s
- Energy of valence *p* orbital E_p
- Radius of valence *s* orbital r_s
- Radius of valence *p* orbital r_p

r_s, r_p, $E_s/\sqrt{Z_v}$, $E_p/\sqrt{z_v}$,
for A and B atoms



What's on the axes?

Linear combination of (possibly all) the initial dimensions



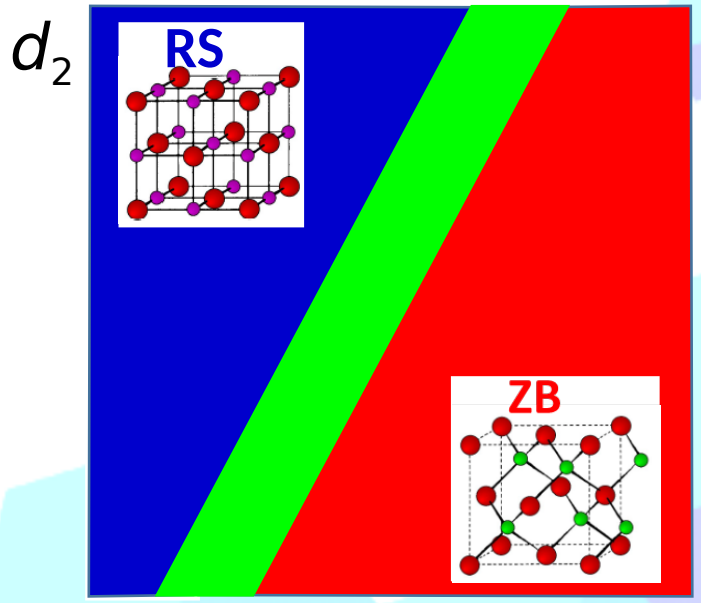
Saad, …, Chelikowsky, and Andreoni, PRB 85, 104104 (2012)

Compressed sensing: the quest for descriptors and predictive models

82 octet AB binary compounds

Ansatz: atomic features

- HOMO
- LUMO
- Ionization Potential
- Electron Affinity
- Radius of valence *s* orbital
- Radius of valence *p* orbital
- Radius of valence *d* orbital
- Thousands to billions of non-linear functions of the above



Rock salt
RS / ZB
Zinc blende

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

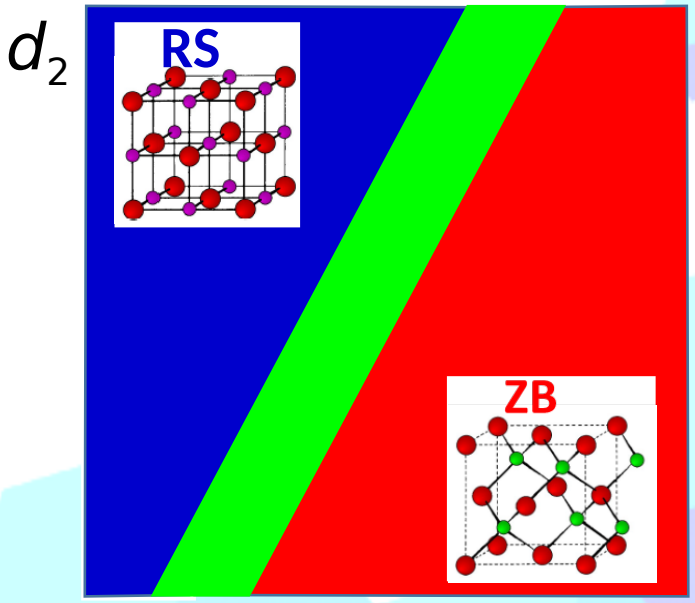
E(Rock salt) – *E*(Zinc blende)

Compressed sensing: the quest for descriptors and predictive models

82 octet AB binary compounds

Ansatz: atomic features

- HOMO
- LUMO
- Ionization Potential
- Electron Affinity
- Radius of valence *s* orbital
- Radius of valence *p* orbital
- Radius of valence *d* orbital
- Thousands to billions of non-linear functions of the above



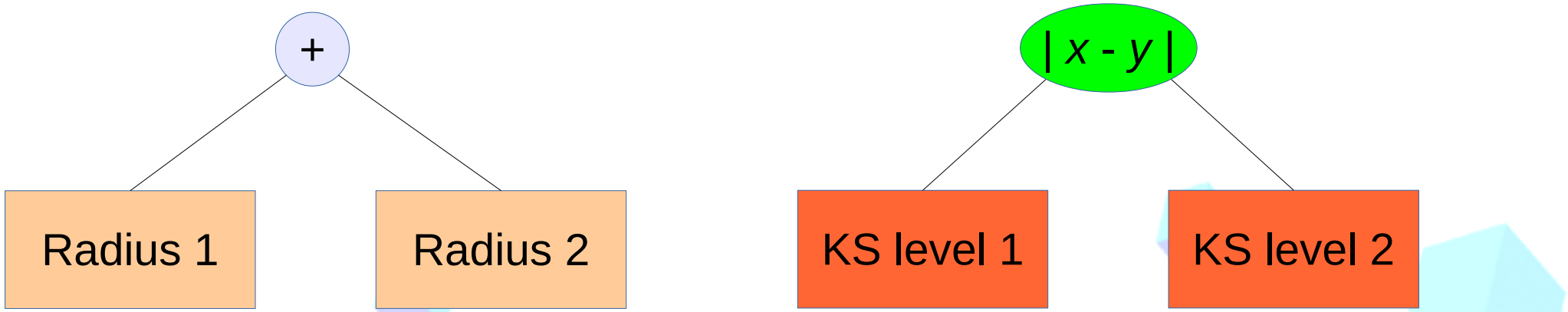
Rock salt
RS / ZB
Zinc blende

Symbolic Regression

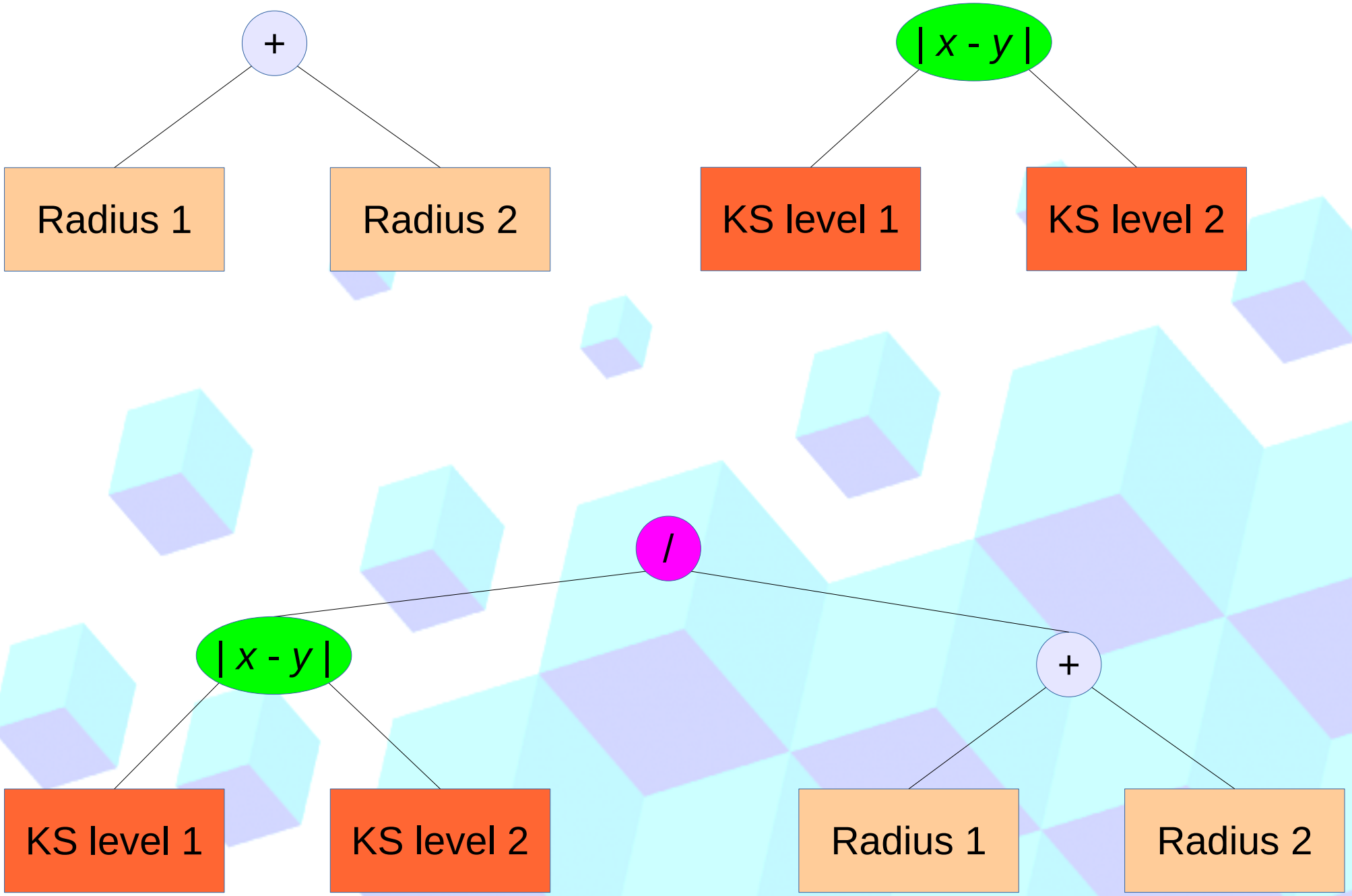
$$P = c_1 d_1 + c_2 d_2 + \ldots c_n d_n$$

E(Rock salt) – *E*(Zinc blende)

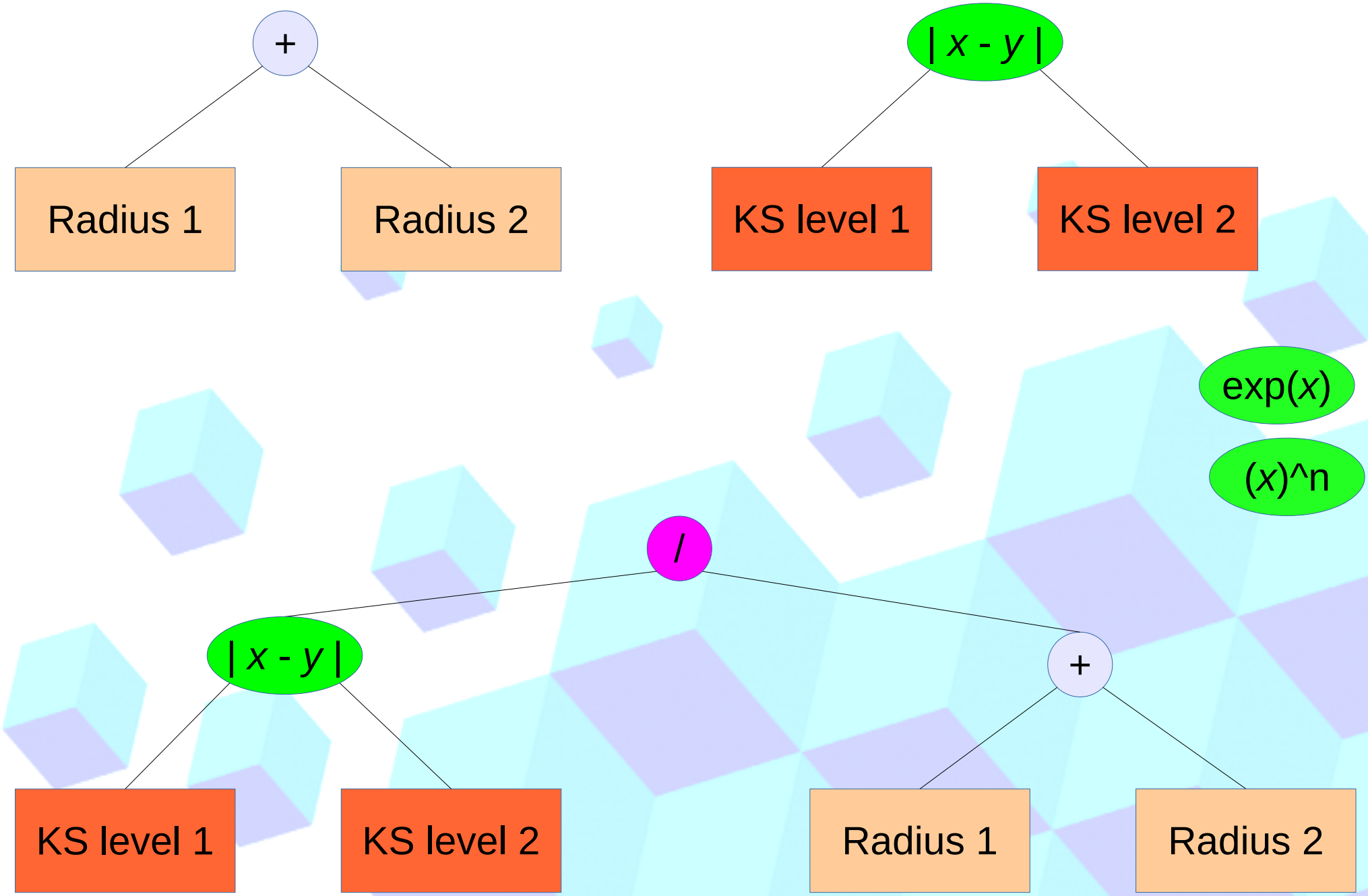
Systematic construction of the feature space



Systematic construction of the feature space



Systematic construction of the feature space



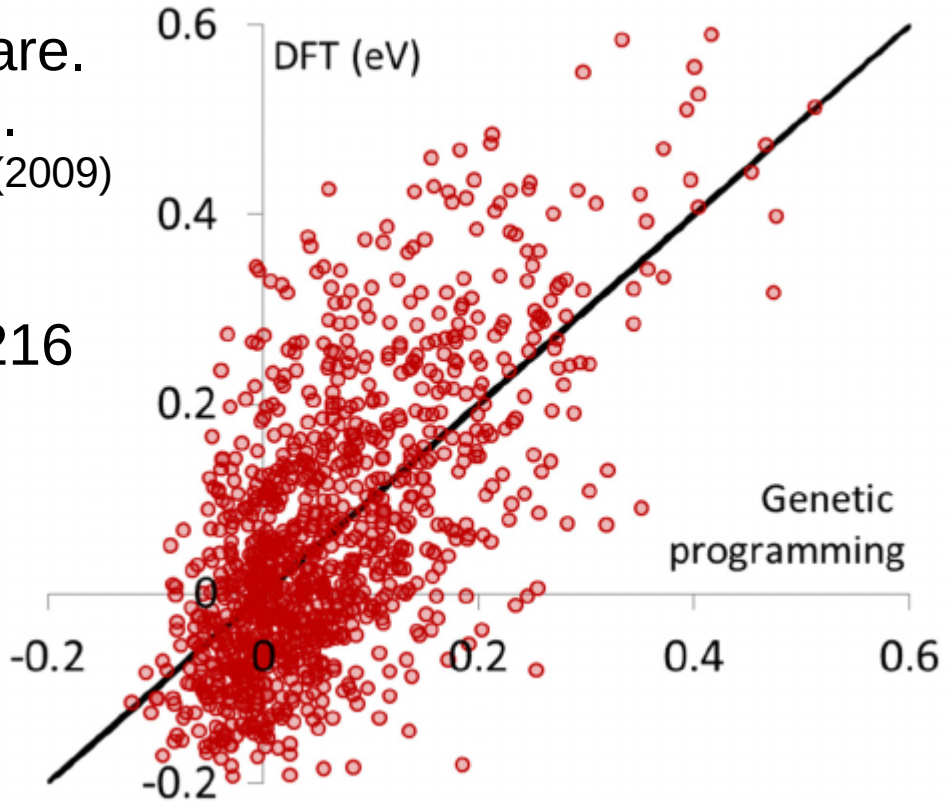
Systematic construction of the feature space: EUREQA

EUREQA: genetic programming software.
Global optimization (genetic algorithm).
Schmidt M., Lipson H., Science, Vol. 324, No. 5923, (2009)

T. Müller et al. PRB **89** 115202 (2014):
Data: ~1000 amorphous structures of 216 Si atoms (saturated)

Property: hole trap depth

$$\frac{\min(1.66355,a)\max(5.37551,c) - f - bd}{g} - h\max(3.42929,e),$$



Descriptor (candidates: 242)

a The largest distance between a H atom and its nearest Si neighbor
b The shortest distance between a Si atom and its sixth-nearest Si neighbor
c The maximum bond valence sum on a Si atom
d The smallest value for the fifth-smallest relative bond length around a Si atom
e The fourth-shortest distance between a Si atom and its eighth-nearest neighbor
f The second-shortest distance between a Si atom and its fifth-nearest neighbor
g The third-shortest distance between a Si atom and its sixth-nearest neighbor
h The H-Si nearest-neighbor distance for the hydrogen atom with the fourth-smallest difference between the distances to the two Si atoms nearest to a H atom

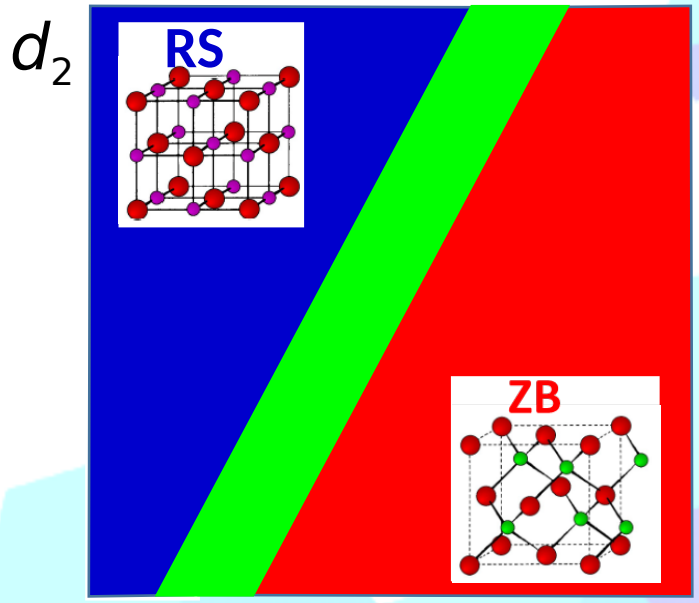
Building block	
Constant value	Exponential
Input variable	Natural logarithm
Addition	Power
Subtraction	Square root
Multiplication	Logistic function
Division	Minimum
Negation	Maximum
	Absolute value

Compressed sensing: the quest for descriptors and predictive models

82 octet AB binary compounds

Ansatz: atomic features

- HOMO
- LUMO
- Ionization Potential
- Electron Affinity
- Radius of valence *s* orbital
- Radius of valence *p* orbital
- Radius of valence *d* orbital
- Thousands of non-linear functions of the above



Rock salt
RS / ZB
Zinc blende

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

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

E(Rock salt) – *E*(Zinc blende)

Compressed sensing: the quest for descriptors and predictive models

Ideal method: regression with ℓ_0 regularization

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

Optimal solution
Non-polinomial complexity
Small # columns in ***D***

Compressed sensing: the quest for descriptors and predictive models

Ideal method: regression with ℓ_0 regularization

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

Optimal solution
Non-polinomial complexity
Small # columns in ***D***

$\|\boldsymbol{c}\|_0$	# of nonzero elements of ***c***
$\|\boldsymbol{c}\|_2$	Euclidean. Square root of sum of squares of the elements of ***c***)

Compressed sensing: the quest for descriptors and predictive models

Ideal method: regression with ℓ_0 regularization

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

Optimal solution
Non-polinomial complexity
Small # columns in ***D***

$\|\boldsymbol{c}\|_0$ # of nonzero elements of ***c***

$\|\boldsymbol{c}\|_2$ Euclidean. Square root of sum of squares of the elements of ***c***)

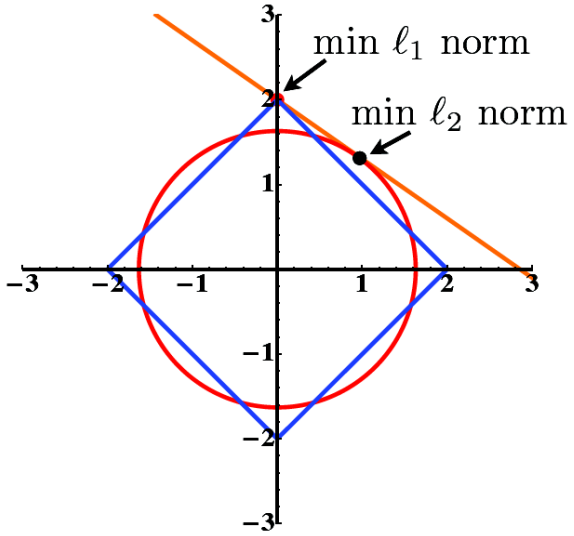
For matrices D with uncorrelated columns: LASSO

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}}\|\boldsymbol{P}-\boldsymbol{D}\boldsymbol{c}\|_2^2+\lambda\|\boldsymbol{c}\|_1$$

(Possibly) optimal solution
Convex optimization
Moderate # columns in ***D***

$\|\boldsymbol{c}\|_1$ “Manhattan”. Sum of absolute values of the elements of ***c***

Compressed sensing: the quest for descriptors and predictive models



$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_1$$

(Possibly) optimal solution
Convex optimization
Moderate # columns in ***D***

$\|\boldsymbol{c}\|_1$ "Manhattan". Sum of absolute values of the elements of ***c***

Compressed sensing in materials science

PRL 113, 185501 (2014) PHYSICAL REVIEW LETTERS week ending 31 OCTOBER 2014

Lattice Anharmonicity and Thermal Conductivity from Compressive Sensing of First-Principles Calculations

Fei Zhou (周非)
Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California 94550, USA

Weston Nielson, Yi Xia, and Vidvuds Ozoliņš
Department of Materials Science and Engineering, University of California, Los Angeles, California 90095-1595, USA
(Received 22 April 2014; published 27 October 2014)

Compressed modes for variational problems in mathematics and physics

Vidvuds Ozoliņš[a,1], Rongjie Lai[b,1], Russel Caflisch[c,1], and Stanley Osher[c,1,2]

Departments of [a]Materials Science and Engineering, and [c]Mathematics, University of California, Los Angeles, CA 90095-1555; and [b]Department of Mathematics, University of California, Irvine, CA 92697-3875

Contributed by Stanley Osher, October 8, 2013 (sent for review September 3, 2013)

PHYSICAL REVIEW B **87**, 035125 (2013)

Compressive sensing as a paradigm for building physics models

Lance J. Nelson and Gus L. W. Hart
Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602, USA

Fei Zhou (周非) and Vidvuds Ozoliņš*
Department of Materials Science and Engineering, University of California, Los Angeles, California 90095, USA
(Received 26 June 2012; revised manuscript received 26 September 2012; published 18 January 2013)

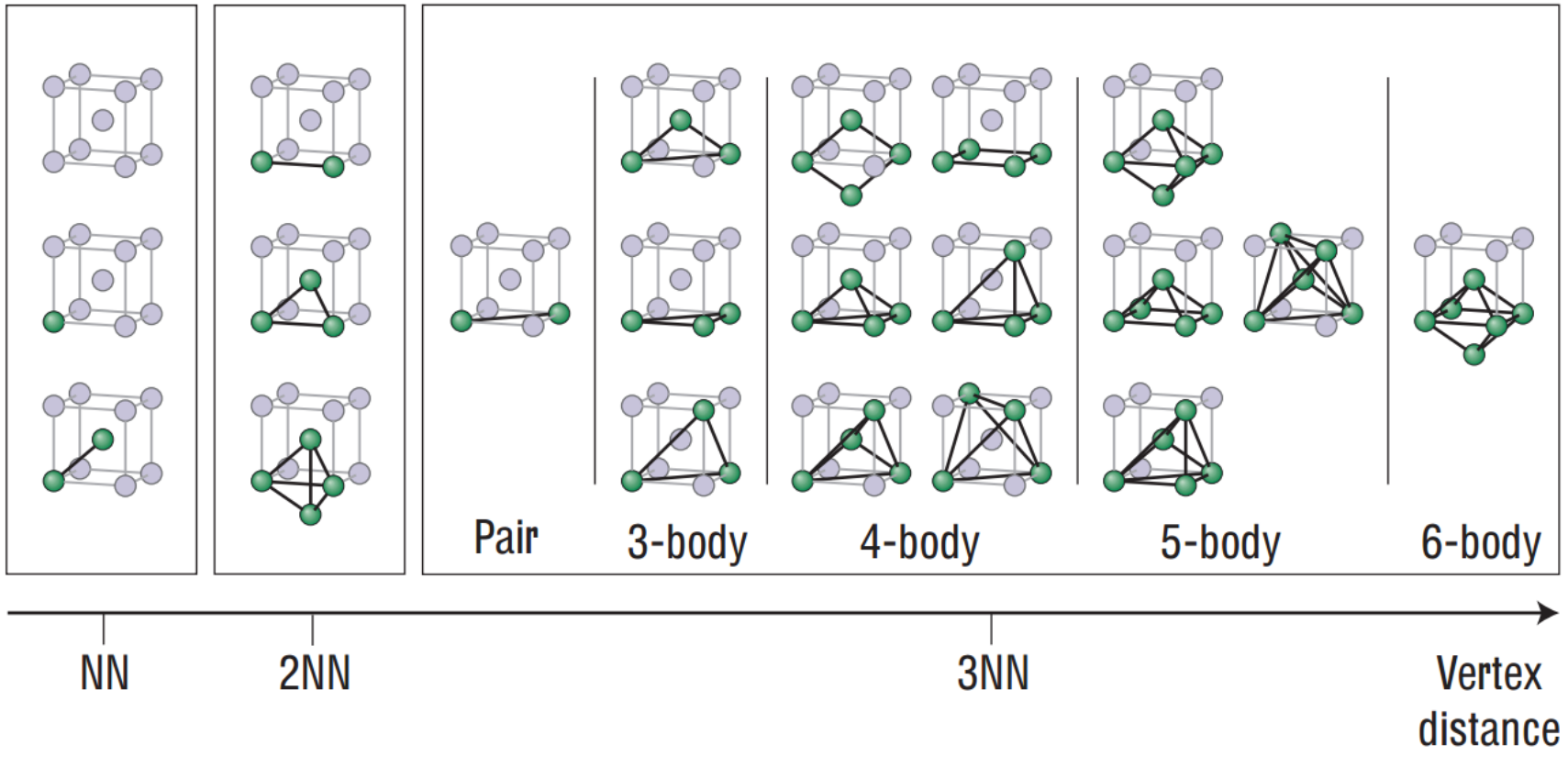
PHYSICAL REVIEW B **87**, 035125 (2013)

Compressive sensing as a paradigm for building physics models

Lance J. Nelson and Gus L. W. Hart
Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602, USA

Fei Zhou (周非) and Vidvuds Ozoliņš*
Department of Materials Science and Engineering, University of California, Los Angeles, California 90095, USA
(Received 26 June 2012; revised manuscript received 26 September 2012; published 18 January 2013)

$$E(\sigma) = E_0 + \sum_f \bar{\Pi}_f(\sigma) J_f$$

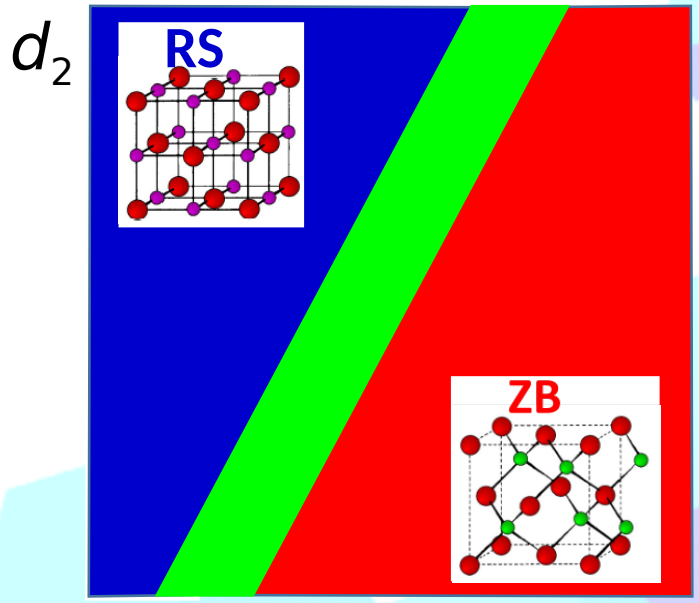


Compressed sensing: the quest for descriptors and predictive models

82 octet AB binary compounds

Ansatz: atomic features

- HOMO
- LUMO
- Ionization Potential
- Electron Affinity
- Radius of valence *s* orbital
- Radius of valence *p* orbital
- Radius of valence *d* orbital
- Billions of non-linear functions of the above



Rock salt
RS / ZB
Zinc blende

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

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

E(Rock salt) – *E*(Zinc blende)

Compressed sensing: the quest for descriptors and predictive models

From orthogonal matching pursuit ….

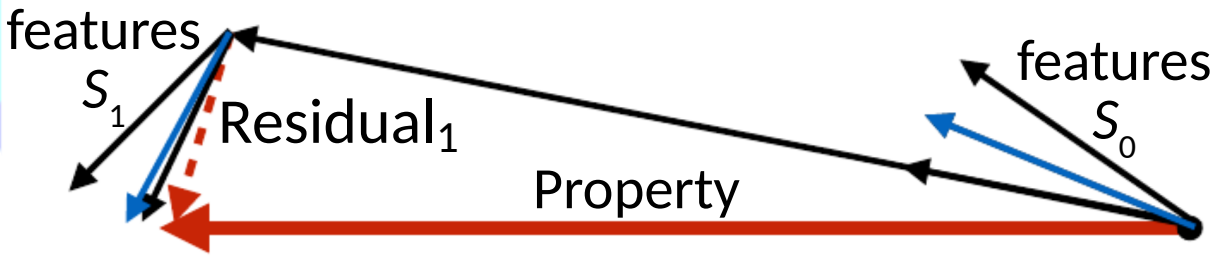
Local-optimum solution
Huge # columns in D

Compressed sensing: the quest for descriptors and predictive models

From orthogonal matching pursuit ….

Local-optimum solution
Huge # columns in D

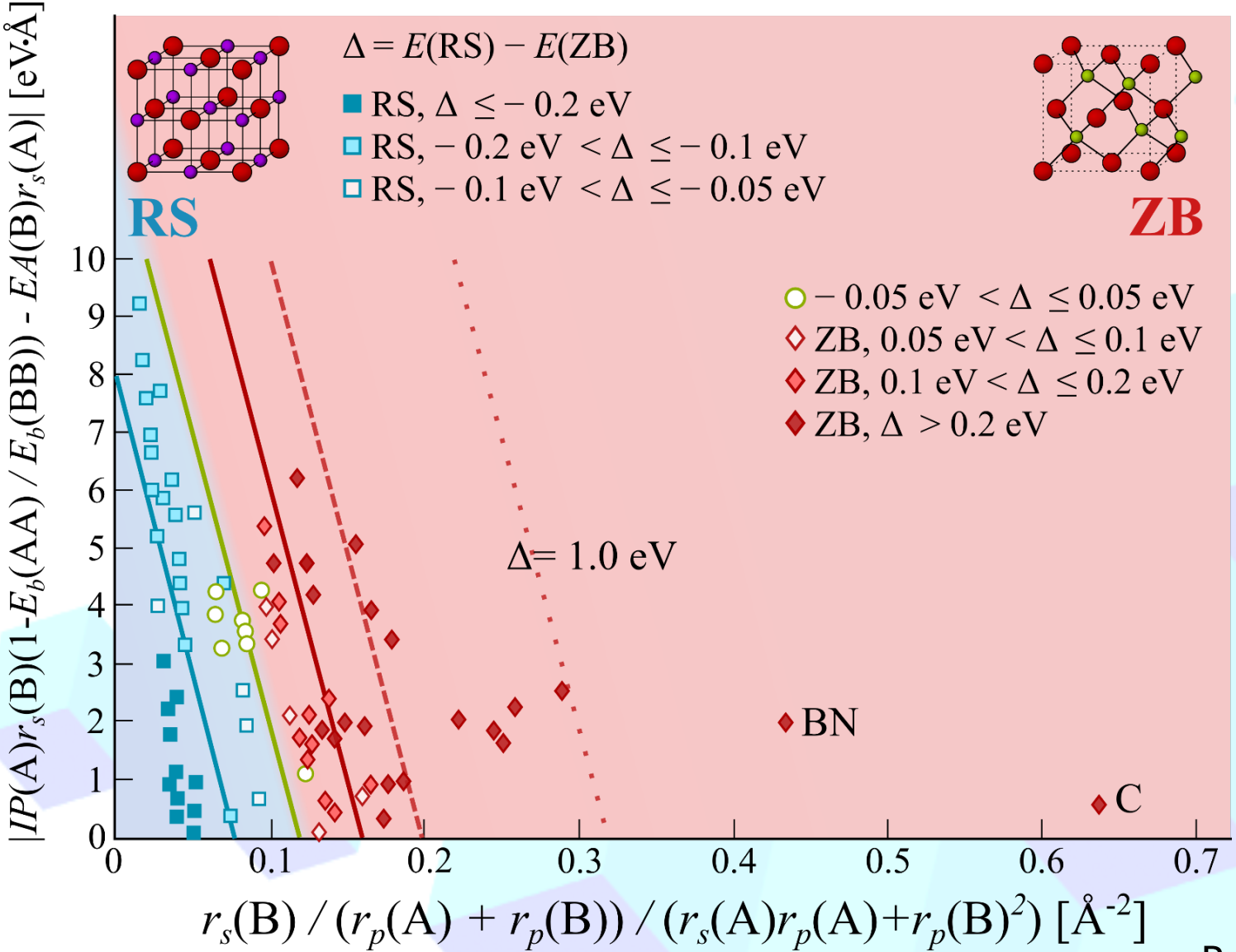
… to Sure Independence Screening + Sparsifying Operator (SISSO)



Proxy of
global-optimum solution
Huge # columns in D

R. Ouyang *et al.* PRM **2**, 083802 (2018), published 7 August 2018)

Compressed sensing: the quest for descriptors and predictive models



Structure map with SISSO, starting from 7 atomic + 6 dimer features
Feature space: 10^{11} features

R. Ouyang *et al.* PRM (2018)

Compressed sensing: the quest for descriptors and predictive models

$$\arg\min_{\boldsymbol{c}} \left(\|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2 + \lambda \|\boldsymbol{c}\|_0 \right)$$

Compressed-sensing-based model identification:
Shares concepts with

Regularized regression. But: Massive sparsification.

Dimensionality reduction. But supervised, and yielding sparse, “inspectable” descriptors

Feature/Basis-set selection/extraction. But: non-greedy solver.

Symbolic regression. But: deterministic solver.

One descriptor to rule them all: Multi-task learning

$$\{P^{(1)}, P^{(2)}, \ldots, P^{N^{\mathrm{T}}}\}$$

One descriptor to rule them all: Multi-task learning

$$\{P^{(1)}, P^{(2)}, \ldots, P^{N^{\mathrm{T}}}\} \longrightarrow P^k = \boldsymbol{d} \cdot \boldsymbol{c}^k$$

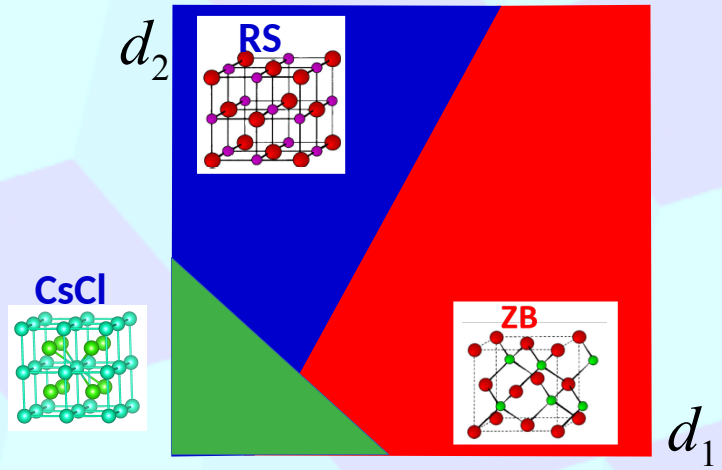
One descriptor to rule them all: Multi-task learning

$$\{P^{(1)}, P^{(2)}, \ldots, P^{N^{\mathrm{T}}}\} \longrightarrow P^k = \boldsymbol{d} \cdot \boldsymbol{c}^k$$

One descriptor to rule them all: Multi-task learning

$$\{P^{(1)}, P^{(2)}, \ldots, P^{N^{\mathrm{T}}}\} \longrightarrow P^k = \boldsymbol{d} \cdot \boldsymbol{c}^k$$

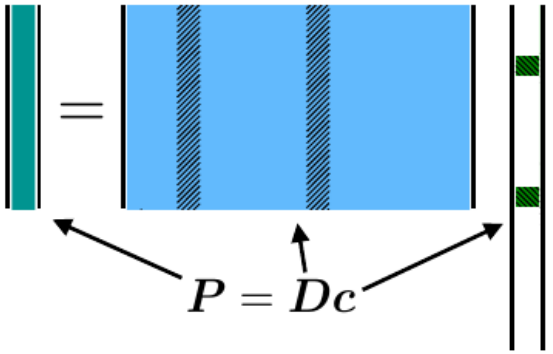
Application: multi-phase stability diagram
Properties: crystal-structure formation energies



One descriptor to rule them all: Multi-task learning

$$\{P^{(1)}, P^{(2)}, \ldots, P^{N_T}\} \Rightarrow P^k = \boldsymbol{d} \cdot \boldsymbol{c}^k$$

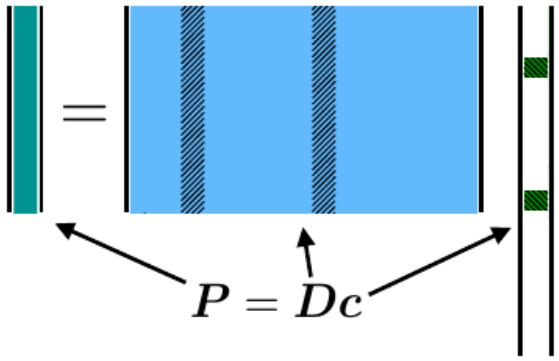
ST-SISSO



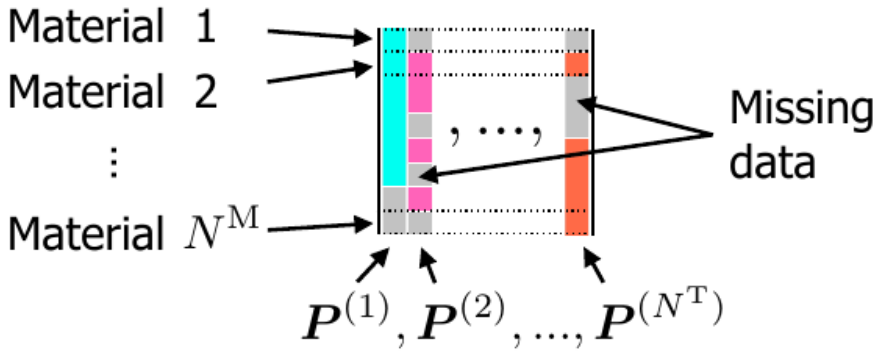
One descriptor to rule them all: Multi-task learning

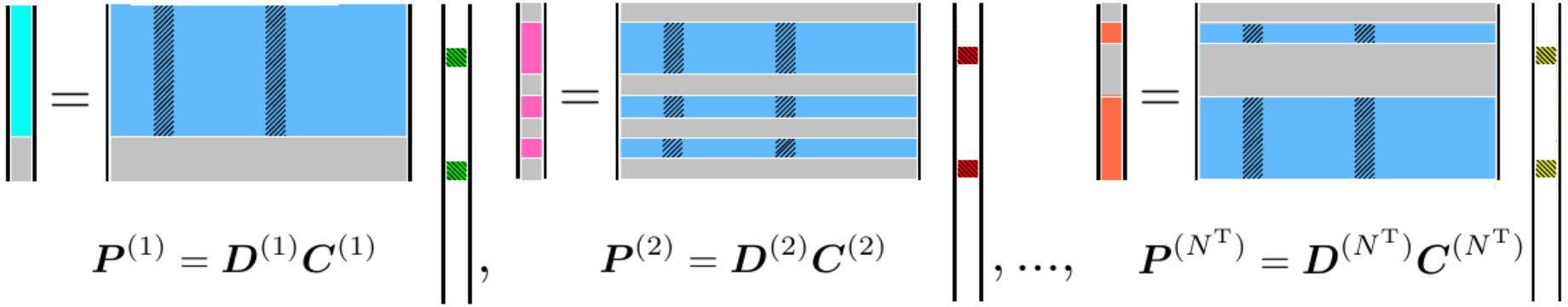
$$\{P^{(1)}, P^{(2)}, \ldots, P^{N^{\mathrm{T}}}\} \Rightarrow P^k = \boldsymbol{d} \cdot \boldsymbol{c}^k$$

ST-SISSO

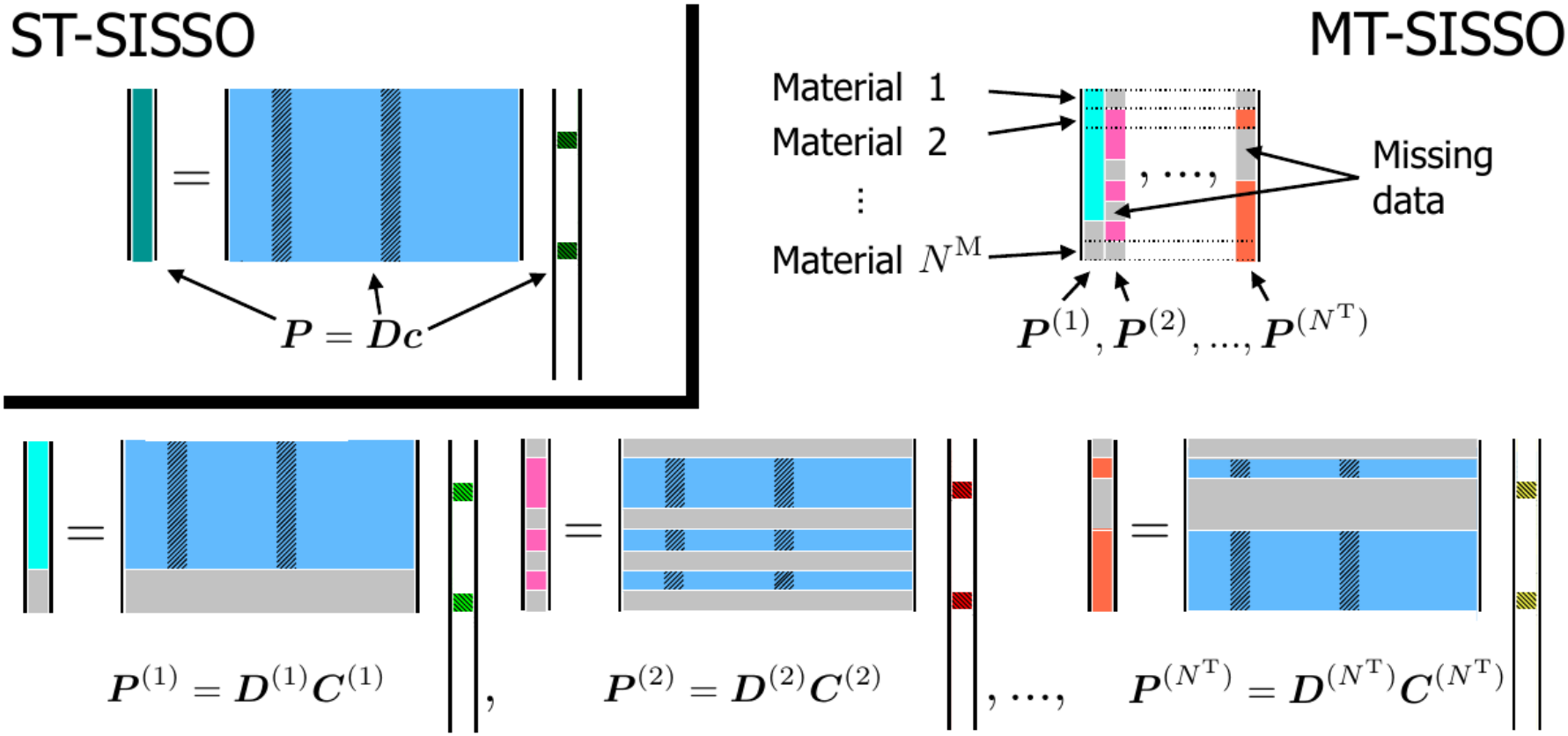


MT-SISSO

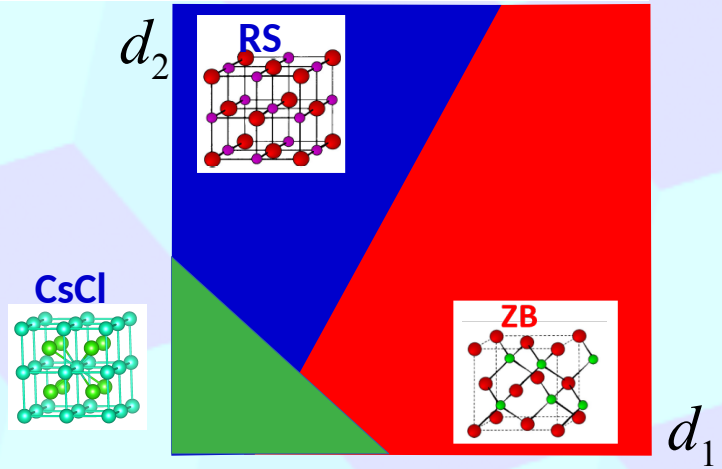




One descriptor to rule them all: Multi-task learning



Application: multi-phase stability diagram
Properties: crystal-structure formation energies



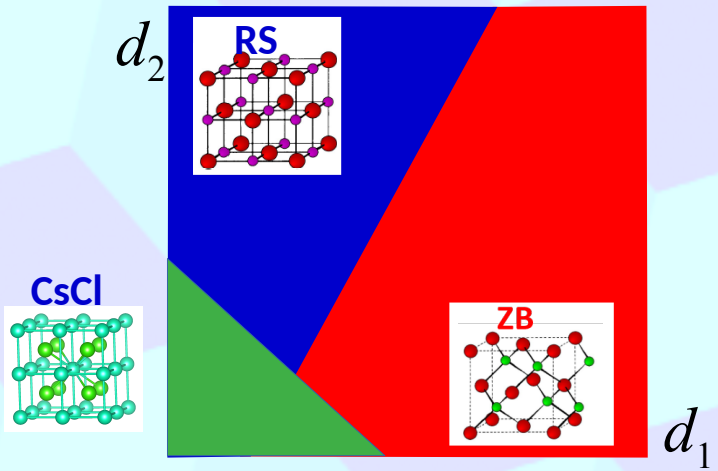
One descriptor to rule them all: Multi-task learning

$$\{P^{(1)}, P^{(2)}, \ldots, P^{N^{\mathrm{T}}}\} \longrightarrow P^k = \boldsymbol{d} \cdot \boldsymbol{c}^k$$

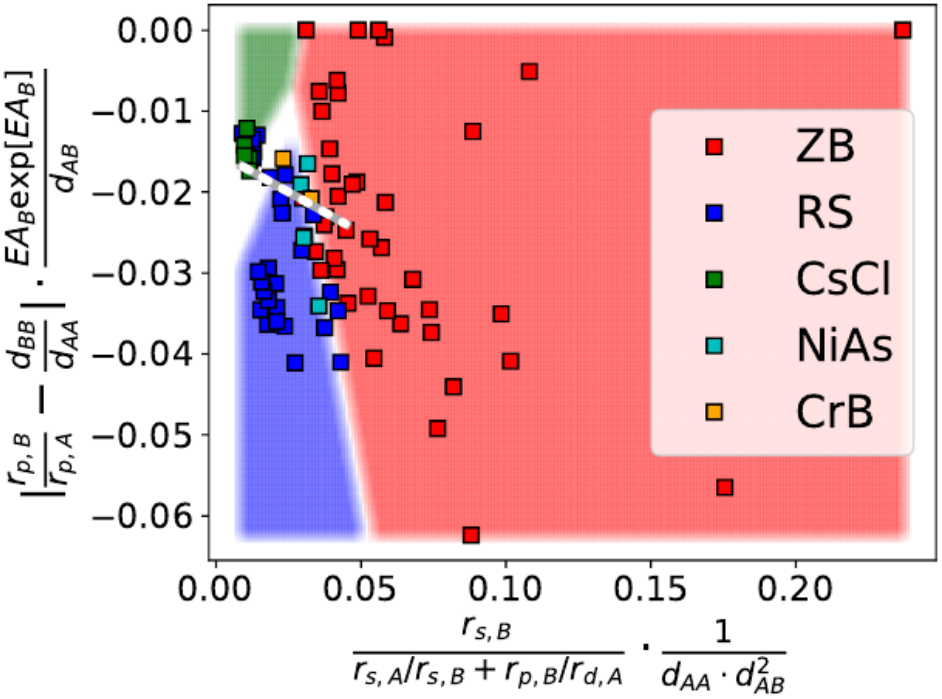
$$\arg\min_{\boldsymbol{c}} \left(\|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2 + \lambda \|\boldsymbol{c}\|_0 \right)$$

$$\arg\min_{\boldsymbol{C}} \sum_{k=1}^{N^{\mathrm{T}}} \frac{1}{N_k^{\mathrm{M}}} \left\| \boldsymbol{P}^k - \boldsymbol{D}^k \boldsymbol{C}^k \right\|_2^2 + \lambda \|\boldsymbol{C}\|_0$$

Application: multi-phase stability diagram
Properties: crystal-structure formation energies

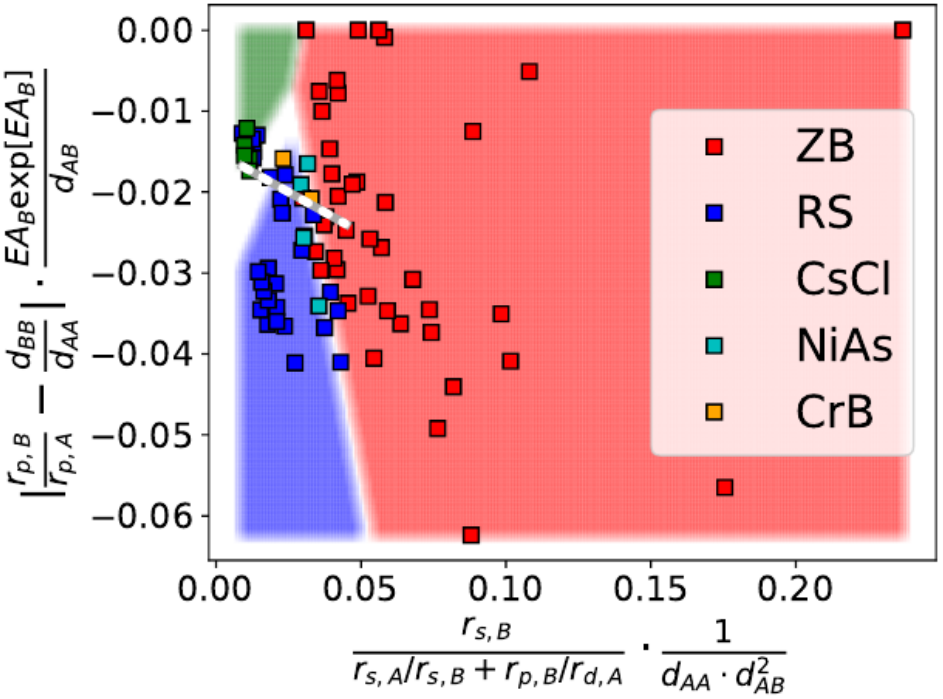


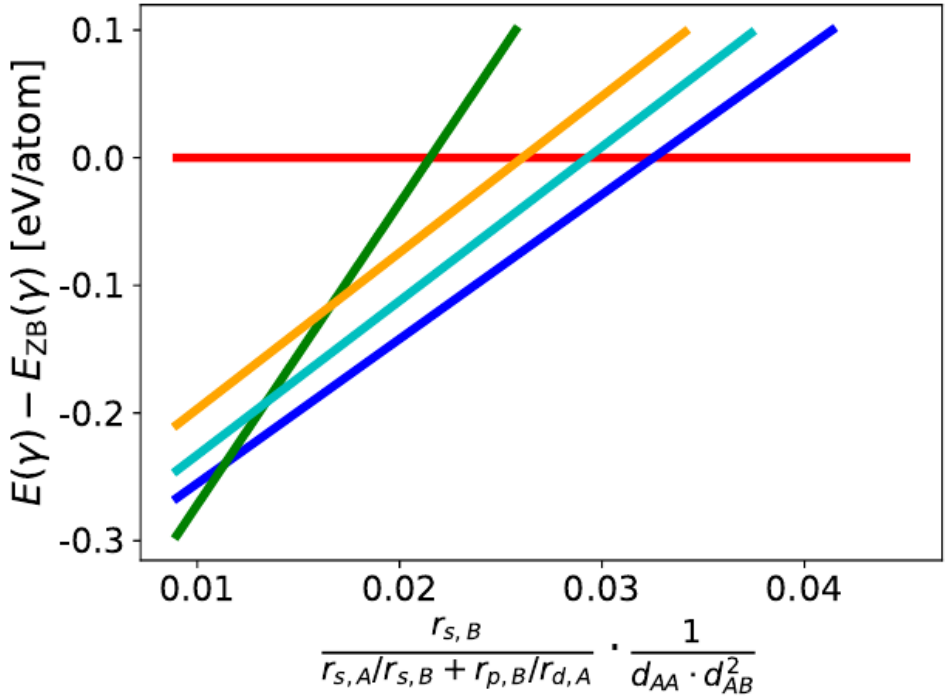
One descriptor to rule them all: Multi-task SISSO
Energy differences among 5 crystal structures.



Ouyang, Ahmetcik, Carbogno, Scheffler & LMG, J. Phys. Mater. 2, 024002 (2019)

One descriptor to rule them all: Multi-task SISSO
Energy differences among 5 crystal structures.





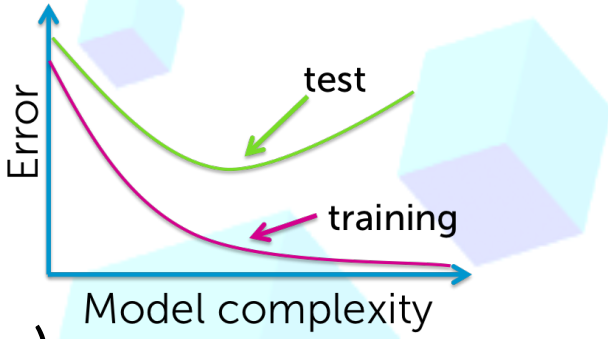
Ouyang, Ahmetcik, Carbogno, Scheffler & LMG, J. Phys. Mater. 2, 024002 (2019)

A general scheme about training, cross-validation, test

Training: input (features, descriptor) + labels (values target property) → yields one model which minimizes a cost function (incl. regularization)

A general scheme about training, cross-validation, test

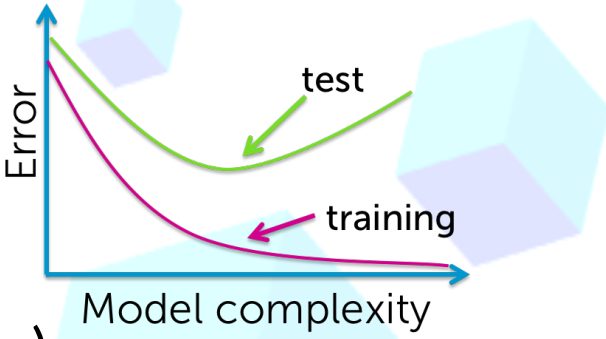
Training: input (features, descriptor) + labels (values target property)
→ yields one model which minimizes a cost function (incl. regularization)



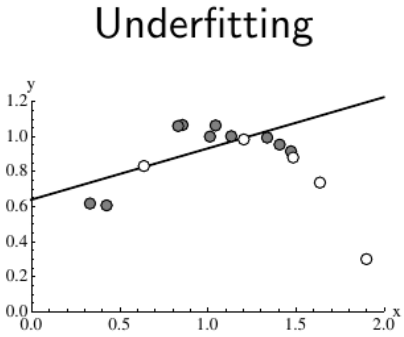
Cross-validation: used to tune model-complexity
- perform training *n* times on different split of data.
  Training + test/validation sets
→ yields one model that minimizes the test (validation) error

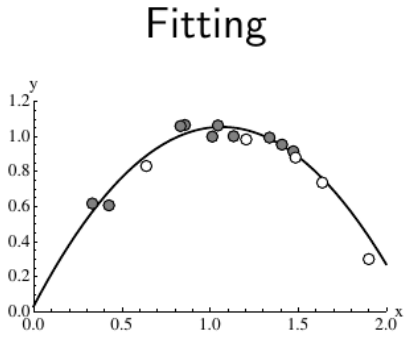
A general scheme about training, cross-validation, test

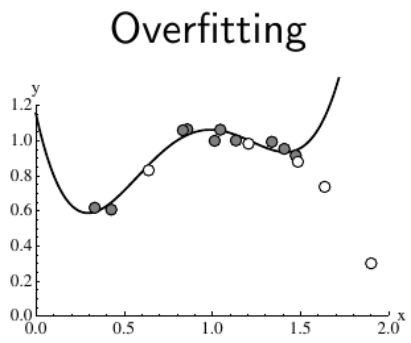
Training: input (features, descriptor) + labels (values target property)
→ yields one model which minimizes a cost function (incl. regularization)



Cross-validation: used to tune model-complexity
- perform training *n* times on different split of data.
  Training + test/validation sets
→ yields one model that minimizes the test (validation) error

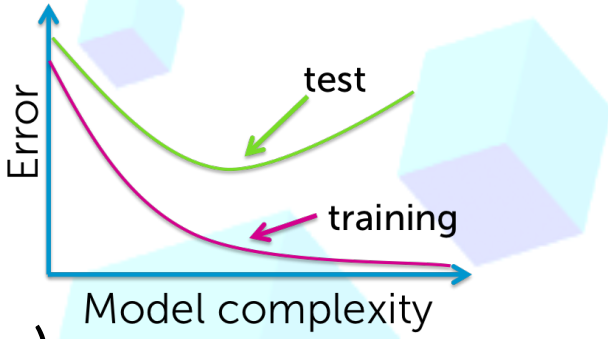




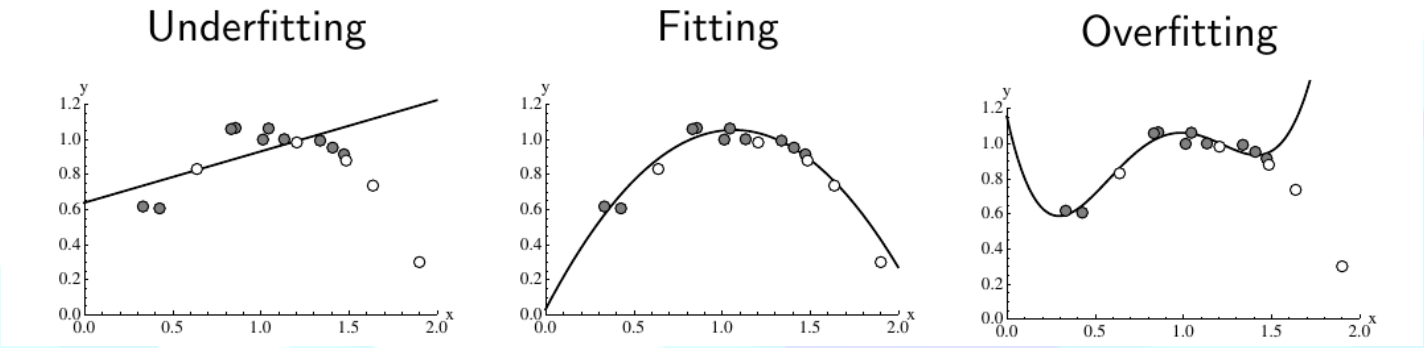


A general scheme about training, cross-validation, test

Training: input (features, descriptor) + labels (values target property)
→ yields one model which minimizes a cost function (incl. regularization)



Cross-validation: used to tune model-complexity
- perform training *n* times on different split of data.
  Training + test/validation sets
→ yields one model that minimizes the test (validation) error



Test: evaluation of the performance of the model on data never used for training (i.e., the whole cross-validation procedure), aka left-out set

Data-driven model complexity

- In compressed sensing the “hyperparameters” are
  - the level of sparsity (optimal dimensionality of the model)
  - the size of the feature space

Data-driven model complexity

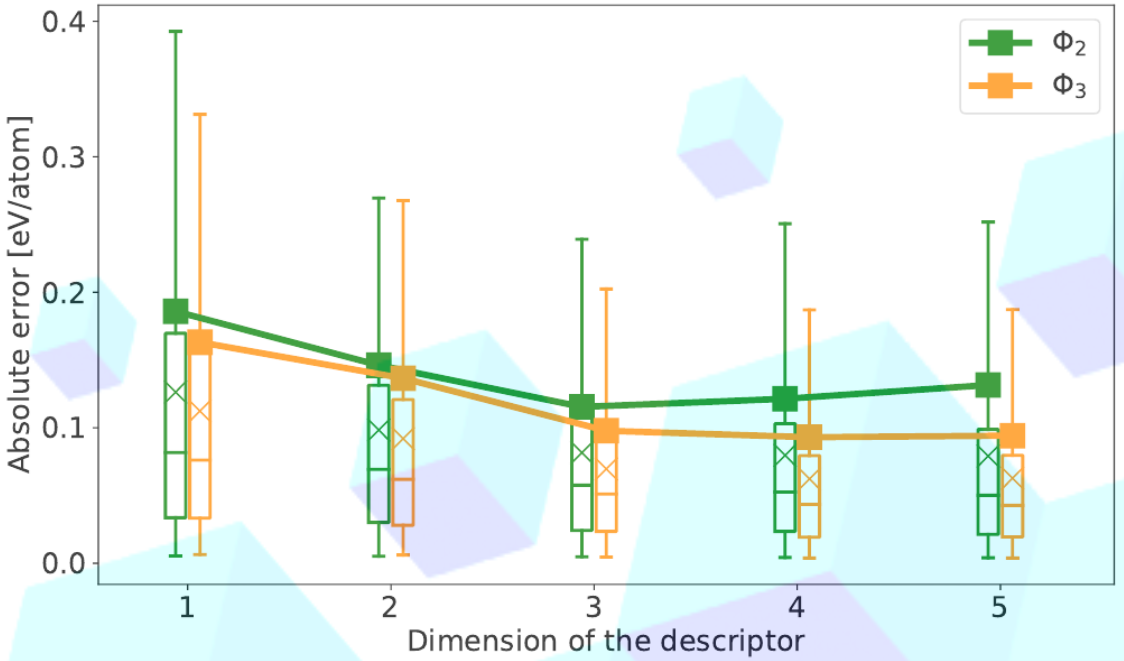
- In compressed sensing the “hyperparameters” are
  - the level of sparsity (optimal dimensionality of the model)
  - the size of the feature space

- Tuned via cross-validation: Iterated random selection of a subset of the data for training and test on the left out set

Data-driven model complexity

- In compressed sensing the “hyperparameters” are
  - the level of sparsity (optimal dimensionality of the model)
  - the size of the feature space

- Tuned via cross-validation: Iterated random selection of a subset of the data for training and test on the left out set



Ouyang, Ahmetcik, Carbogno, Scheffler & LMG, J. Phys. Mater. 2, 024002 (2019)

Charts/maps of materials

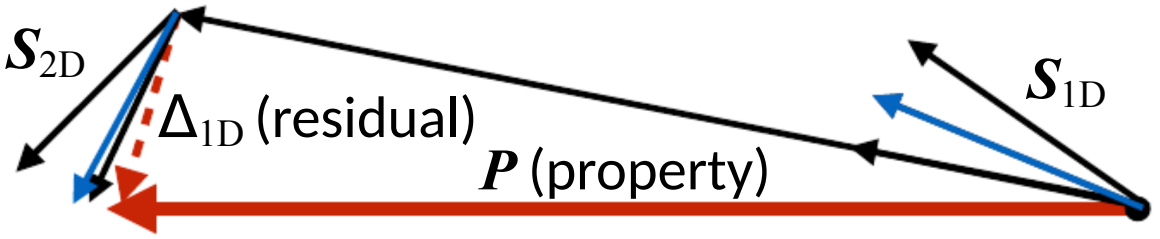
$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}}\left(\|\boldsymbol{P}-\boldsymbol{D}\boldsymbol{c}\|_2^2+\lambda\|\boldsymbol{c}\|_0\right)$$

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

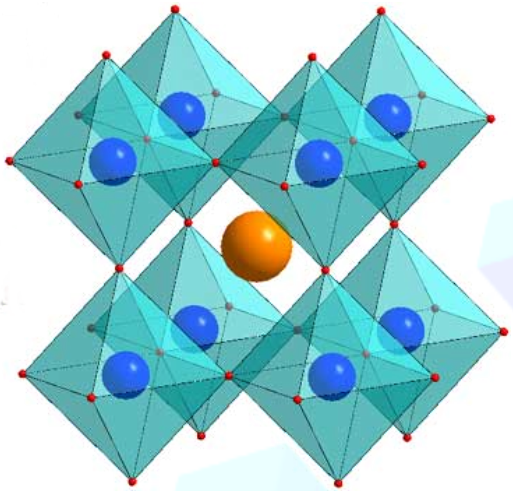
1. # points in the *convex* overlap domain
2. Area of the domain overlap
3. Distance between domains

Good also for multi-categorical problems
(see A. F. Bialon *et al.*, Chem. Mater. **28**, 2550 (2016))

Iterative generation of feature subspaces



Perovskites' stability: an improved Goldschmidt Tolerance Factor

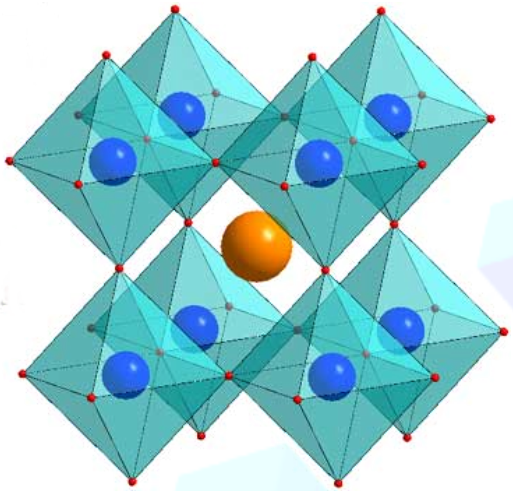


ABX_3

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

Goldschmidt* stable perovskites: 0.825 < ***t*** < 1.059, accuracy 79%

Perovskites' stability: an improved Goldschmidt Tolerance Factor



ABX_3

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

Ionic radius

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

Oxidation state

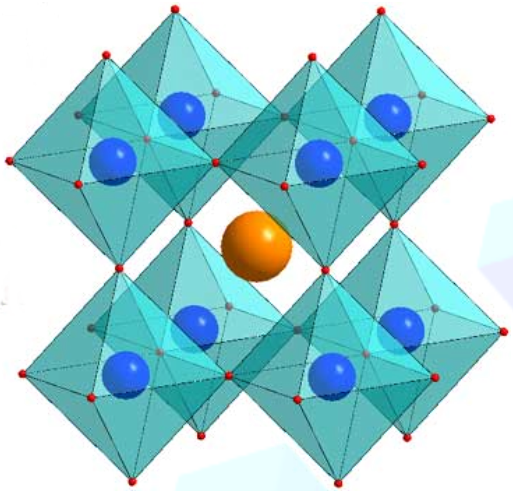
1 / μ = Octahedral factor

Goldschmidt* stable perovskites: 0.825 < ***t*** < 1.059, accuracy 79%

Our stable perovskites: ***τ*** < 4.18, accuracy 92%

Bartel, Sutton, Goldsmith, Ouyang, Musgrave, LMG &Scheffler, Sci. Adv. *5*, *eaav0693 (2019)*

Perovskites' stability: an improved Goldschmidt Tolerance Factor



ABX_3

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

Ionic radius

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

Oxidation state

1 / μ = Octahedral factor

Goldschmidt* stable perovskites: 0.825 < ***t*** < 1.059, accuracy 79%

Our stable perovskites: ***τ*** < 4.18, accuracy 92%

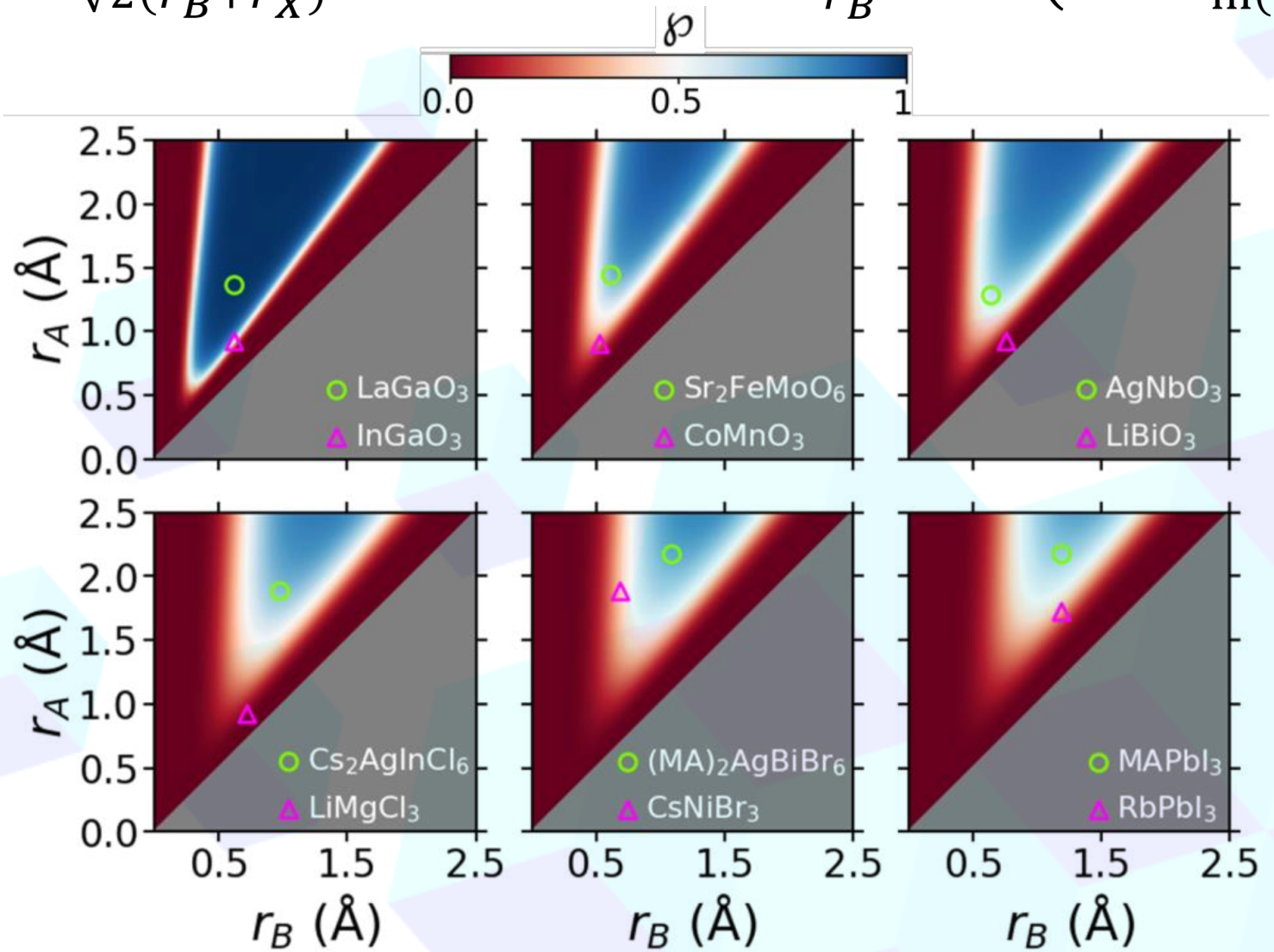
τ < 3.31 or τ > 5.92, 99% accuracy (1/3 of the training data)

τ < 3.31 or τ > 12.08, 100% accuracy (1/4 of the training data)

Bartel, Sutton, Goldsmith, Ouyang, Musgrave, LMG &Scheffler, Sci. Adv. *5*, *eaav0693 (2019)*

Improved Goldschmidt Tolerance Factor: Materials design

$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)} \longrightarrow \tau = \frac{r_X}{r_B} - n_A\left(n_A - \frac{r_A/r_B}{\ln(r_A/r_B)}\right)$$



n_A

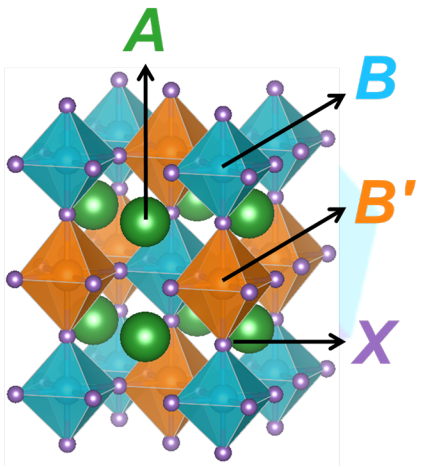
$X = O^{2-}$

$n_A = 1^+$

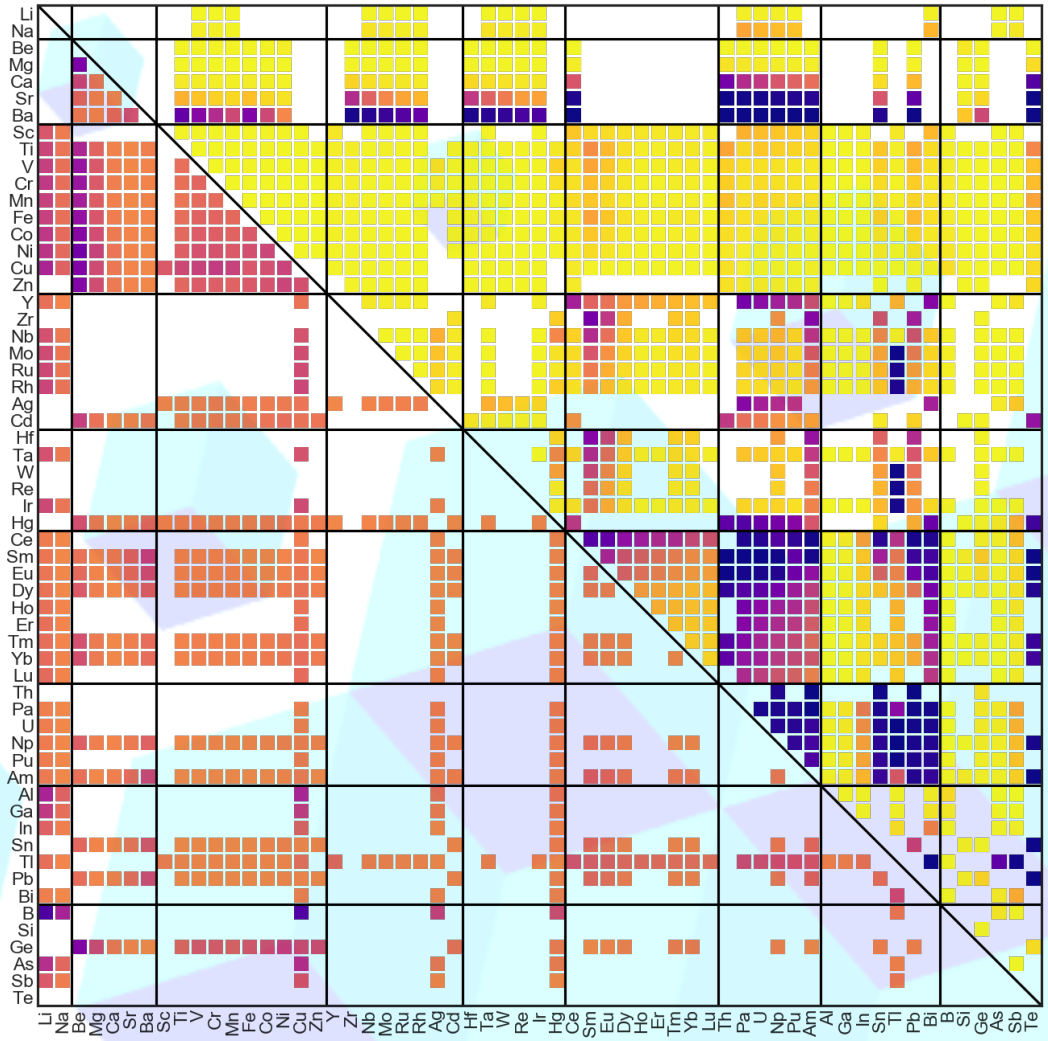
X

Improved Goldschmidt Tolerance Factor: Extension of the materials space

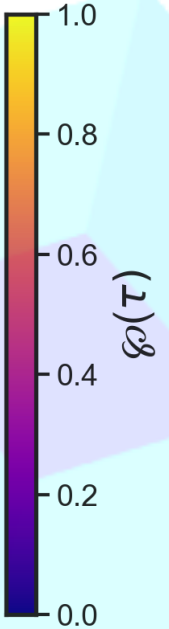
$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)} \longrightarrow \tau = \frac{r_X}{r_B} - n_A\left(n_A - \frac{r_A/r_B}{\ln(r_A/r_B)}\right)$$



$Cs_2BB'Cl_6$

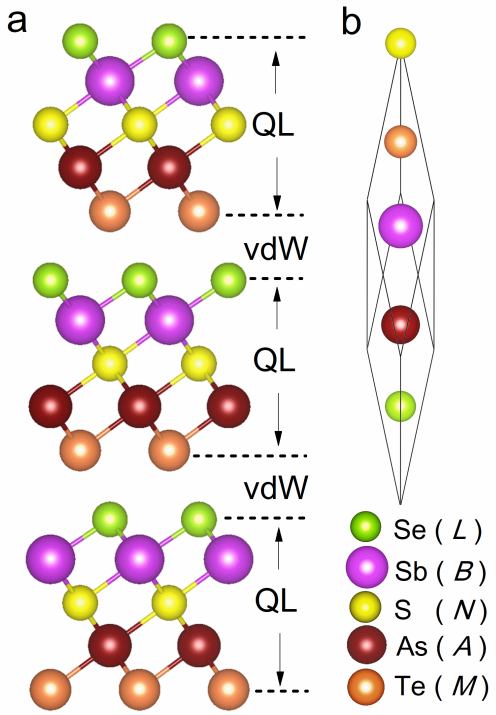


$La_2BB'O_6$



SISSO: predicting new tetradymite topological insulators

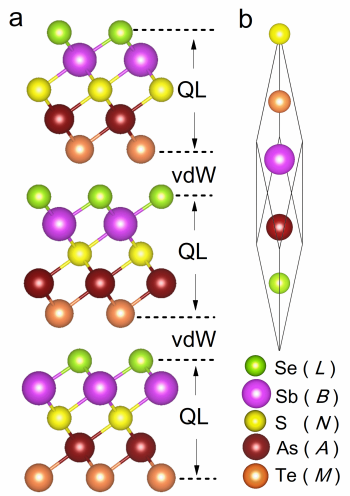
Prototype formula:
AB-LNM
AB = {As,Sb,Bi}
LNM = {S, Se,Te}



Cao, Liu, Ouyang, LMG, Zhou, Scheffler, Zhang, Carbogno, submitted (2019)

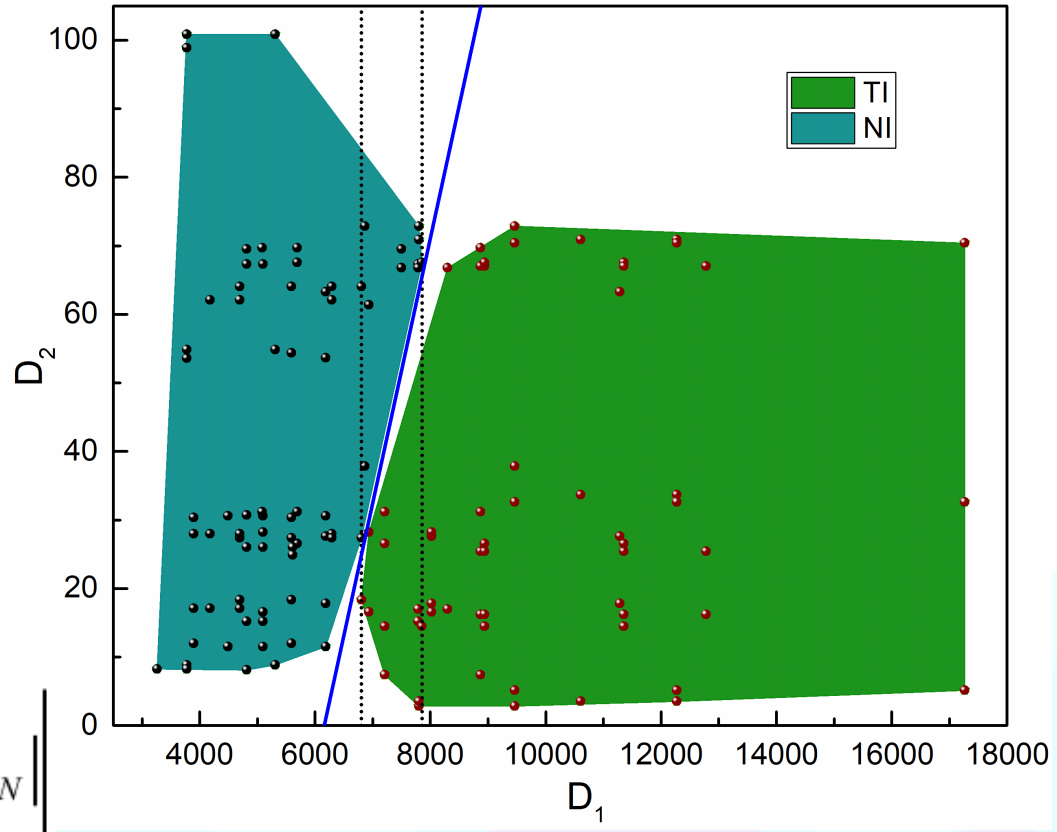
SISSO: predicting new tetradymite topological insulators

Prototype formula:
AB-LNM
AB = {As,Sb,Bi}
LNM = {S, Se,Te}



$$D_1 = (Z_A + Z_B) \cdot (Z_L + Z_M) - |Z_A Z_M - Z_B Z_L|$$

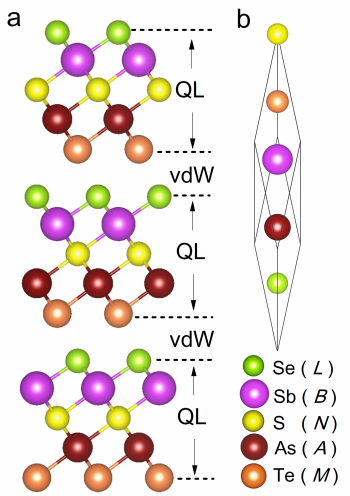
$$D_2 = \left| \frac{(\chi_M + \chi_N) \cdot Z_E}{\chi_A} - (Z_M + Z_N) - |Z_M - Z_N| \right|$$



Cao, Liu, Ouyang, LMG, Zhou, Scheffler, Zhang, Carbogno, submitted (2019)

SISSO: predicting new tetradymite topological insulators

Prototype formula:
AB-LNM
AB = {As,Sb,Bi}
LNM = {S, Se,Te}



$$As_xSb_yBi_{2-x-y}S_aSe_bTe_{3-a-b}$$

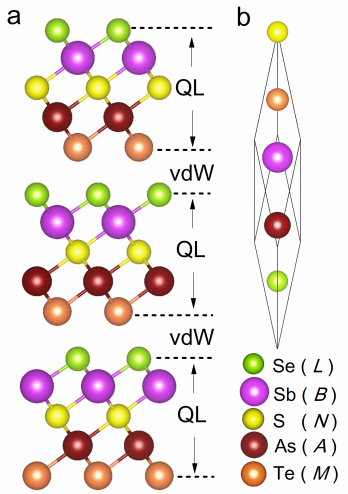
$$D_1 = (Z_A + Z_B)\cdot(Z_L + Z_M) - |Z_AZ_M - Z_BZ_L|$$

$$D_2 = \left|\frac{(\chi_M + \chi_N)\cdot Z_E}{\chi_A} - (Z_M + Z_N) - |Z_M - Z_N|\right|$$

Cao, Liu, Ouyang, LMG, Zhou, Scheffler, Zhang, Carbogno, submitted (2019)

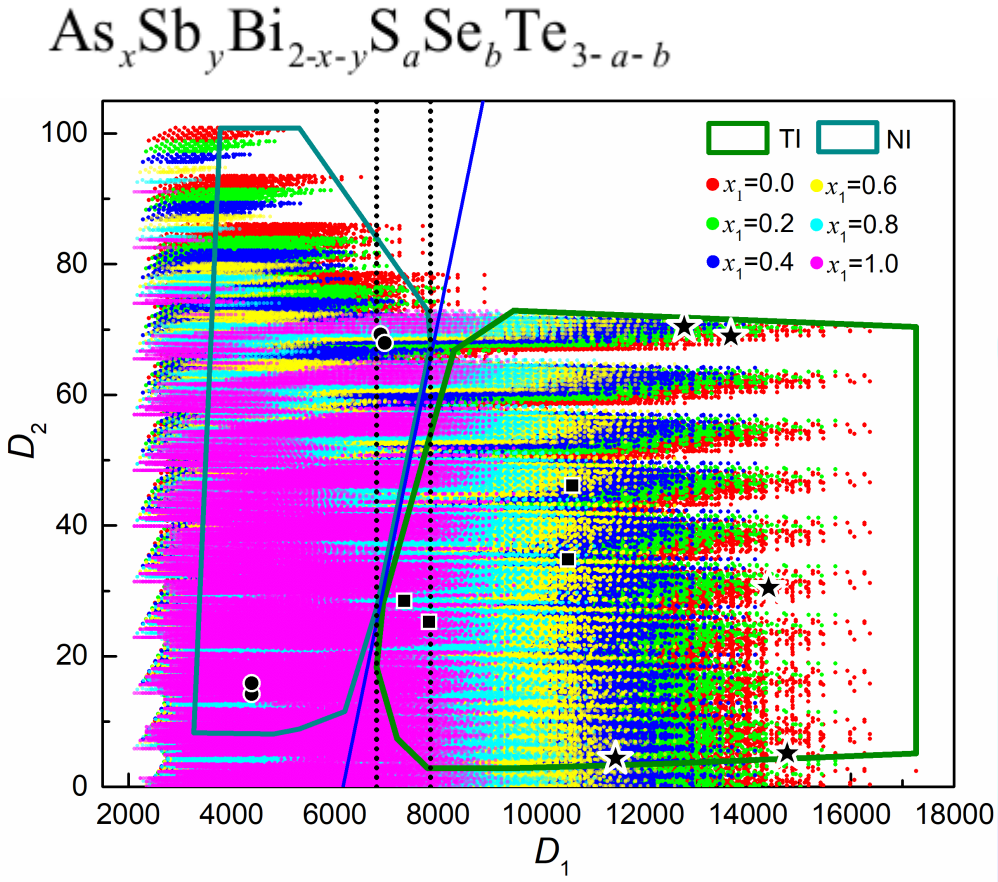
SISSO: predicting new tetradymite topological insulators

Prototype formula:
AB-LNM
AB = {As,Sb,Bi}
LNM = {S, Se,Te}



$$D_1 = (Z_A + Z_B) \cdot (Z_L + Z_M) - |Z_A Z_M - Z_B Z_L|$$

$$D_2 = \left| \frac{(\chi_M + \chi_N) \cdot Z_E}{\chi_A} - (Z_M + Z_N) - |Z_M - Z_N \right.$$



Cao, Liu, Ouyang, LMG, Zhou, Scheffler, Zhang, Carbogno, submitted (2019)

Acknowledgements

Compressed sensing, SISSO, multi-task SISSO, and metal/insulator proof of concepts
Jan Vybiral, Runhai Ouyang, Emre Ahmetcik, Stefano Curtarolo, Christian Carbogno, Sergey Levchenko, Claudia Draxl

Application of SISSO to perovskites
Christopher J. Bartel, Christopher Sutton, Bryan R. Goldsmith, Runhai Ouyang, Charles B. Musgrave

Transparent conducting oxide: NOMAD-kaggle competition
Christopher Sutton, Angelo Ziletti, Takenori Yamamoto, Claudia Draxl, Daan Frenkel, Kristian Thygesen, Samuel Kaski, Bernhard Schölkopf

All the above
Matthias Scheffler

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