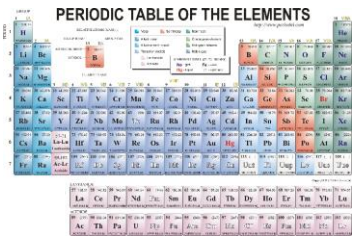
(Big) Data of Materials Science from First Principles -- Critical Role of the Descriptor --

Matthias Scheffler (*)
Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin; http://th.fhi-berlin.mpg.de/

From *the periodic table of the elements* to *a chart of materials*:
Organize materials according to their properties and functions

- figure of merit of thermoelectrics (as function of T)
- turn-over frequency of catalytic materials (as function of T and p)
- efficiency of photovoltaic systems
- etc.

Dmitri Mendeleev
(1834-1907)





(*) Work performed in collaboration with **Luca Ghiringhelli, Jan Vybiral, Claudia Draxl,** Sergey Levchenko, Alexandre Tkatchenko, Patrick Rinke, Xinguo Ren, and Igor Ying Zhang

Materials Genome Initiative
for Global Competiveness

Compute the basic properties („genes“) of many (ten or hundred thousand) materials and disseminate that information to the materials community to enable rapid searches of materials properties and help design improved materials.

To help business discover, develop, and deploy new materials twice as fast, we're launching what we call the Materials Genome Initiative. The invention of silicon circuits and lithium ion batteries made computes and iPods and iPads possible, but it took years to get those technologies from the drawing boards to the market place. We can do it faster.
President Obama
Carnegie Mellon University, June 2011

"twice as fast, at a fraction of the cost"

Materials Genome Initiative for Global Competitiveness http://www. whitehouse.gov/ sites/ deault/files/microsites/stp/materials_genome_initiative-final.pdf

The Four *V* of Big Data and an *A*

Data – data – data (analog to Moore's law)
numbers, arrays, figures, movies, …

(so far: most data are not used and even thrown away)

Big-Data Challenge: *"four V"*:

Volume (amount of data),

Variety (heterogeneity of form and meaning of data),

Veracity (uncertainty of data quality),

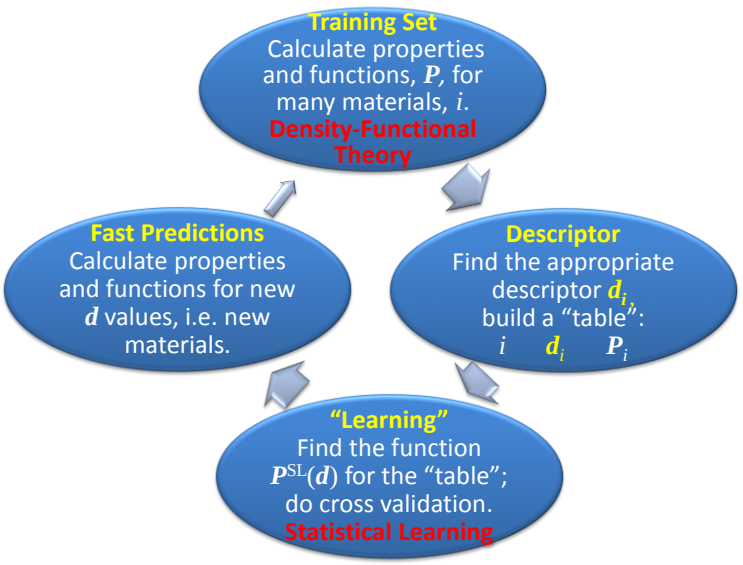
Velocity at which data may change or new data arrive.

Computed data: Query and read out what was stored. (high-throughput screening)
Shouldn't we do more?!

The four V should be complemented by an "***A***", **Big-Data Analytics**:

- identify (so far) hidden trends,
- which materials should be studied next as most promising candidates,
- identify anomalies,
- identify the mechanisms that govern a certain material property or function.

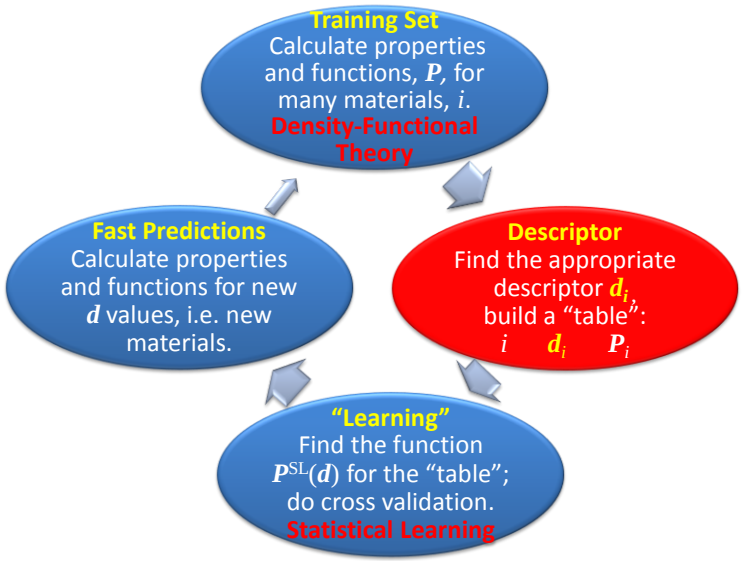
Big-Data Analytics: How to Arrange the Data



$\{Z_I, N_I\}$, T, $\{p\}$ determine the many-body hamiltonian and statistical mechanics

Statistical mechanics does not tell us what the relevant variables are. This is our choice. If we choose well, the results may be useful, if we chose badly, the results (while formally correct) will probably be useless. (Robert Zwanzig)

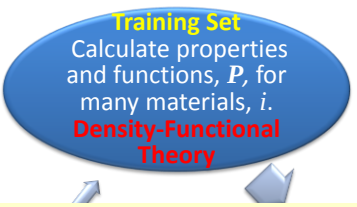
Big-Data Analytics: How to Arrange the Data



$\{Z_I, N_I\}$, T, $\{p\}$ determine the many-body hamiltonian and statistical mechanics

Statistical mechanics does not tell us what the relevant variables are. This is our choice. If we choose well, the results may be useful, if we chose badly, the results (while formally correct) will probably be useless. (Robert Zwanzig)

Big-Data Analytics: How to Arrange the Data



$\{Z_I, N_I\}$, T, $\{p\}$ determine the many-body hamiltonian and statistical mechanics

$\boldsymbol{d}$ characterizes the relevant mechanisms that govern the observed property/function $\boldsymbol{P}$. The $\boldsymbol{d} \rightarrow \boldsymbol{P}^{\mathrm{SL}}$ mapping is complex; identifying the descriptor $\boldsymbol{d}$ from known data $\boldsymbol{P}_i$, is an ill-posed problem (statistical-learning theory): A little error in the data $\boldsymbol{P}_i$ may suggest a different descriptor $\boldsymbol{d}$. Thus, knowledge of the accuracy of data $\boldsymbol{P}_i$ is crucial (veracity). The choice of $\boldsymbol{d}$ is not unique.

A) **Veracity:** Accuracy of state-of-the-art density-functional theory (validation and verification)

B) **Descriptor:** How to find it, how to understand the causality between $\boldsymbol{d}$ and $\boldsymbol{P}^{\mathrm{SL}}$?

http://nomad-repository.eu/cms/ | Welcome to the NoMaD Rep...

The *NoMaD Repository*

by C. Draxl, E. Blokin, L. Ghiringhelli, F. Mohamed, Ch. Carbogno, M. Scheffler

Insight by sharing

Home | NoMaD Team | Why sharing? | Upload | Download | DOIs | Terms and conditions | NoMaD repository | Other repositories

Welcome to the NoMaD Repository

The *NoMaD* (Novel Materials Discovery) *Repository* was established to host, organize, and share materials data.

NoMaD copes with the increasing demand and requirement of storing scientific data and making them available for longer periods. Rules of good scientific practice set by many funding agencies, worldwide, require keeping scientific data for 10 years. ***NoMaD*** offers this for free. ***NoMaD*** also facilitates research groups to share and exchange

News

Currently, the ***NoMaD Repository*** contains **89,464** entries.

Feb. 6, 2015
White House (USA):
"It's Time to Open Materials Science Data" (link)

Open positions

DOIs for datasets can be

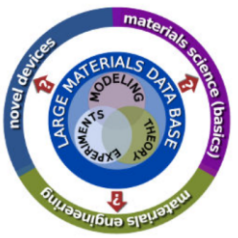
scientific practice set by many funding agencies, worldwide, require keeping scientific data for 10 years. ***NoMaD*** offers this for free. ***NoMaD*** also facilitates research groups to share and exchange their results, inside a single group or between two or more, and to recall what was actually done some years ago.

The ***NoMaD Repository*** enables the confirmatory analysis of materials data, their reuse, and repurposing.

Upload of data is possible without any barrier. Results are accepted in their raw format as produced by the underlying code. The only condition is that the list of authors is provided, and code and code version can be retrieved from the uploaded files. These data can be restricted to the owner or made available to other people (selected by the owner). They can be updated and downloaded at any time.

Read more details concerning the upload. Please, **register** or **login** to participate.

At present, the repository contains *ab initio* electronic-structure data from density-functional theory and methods beyond. At a later stage, it will be extended by force-field studies and by experimental data. We also give an outlook on the ***NoMaD Laboratory*** that will be dedicated to a *Materials Encyclopaedia*, as the basis for complex queries and the development of various data-analytics tools.



Codes: Abinit, crystal, exciting, CASTEP, FHI-aims, Quantum Espresso, VASP – more coming; various xc functionals

Science Data" (link)

Open positions

DOIs for datasets can be requested

Check for related **conferences and workshops**.

We are moving to the HPC Center of the Max Planck Society (RZG). We apologize for any possible instability during the next 2 days.

The ***NoMaD Repository*** is about joining eudat.

Financial Support

Veracity – Validation and Verification

- accuracy of materials-science codes: basis sets, relativity, pseudopotentials, other numerical approximations (verification)
- accuracy of the exchange-correlation functional (validation)

Veracity – Validation and Verification
Comparing Solid State DFT Codes, Basis Sets and Potentials

Code	Version	Basis	Electron treatment	Δ-value	Authors
WIEN2k	13.1	LAPW/APW+lo	all-electron	0 meV/atom	S. Cottenier
FHI-aims	081213	tier2 numerical orbitals	all-electron (relativistic atomic_zora scalar)	0.2 meV/atom	ASE [2]
Exciting	development version	LAPW+xlo	all-electron	0.2 meV/atom	Exciting [10]
FHI-aims	081213	tier2 numerical orbitals	all-electron (relativistic zora scalar 1e-12)	0.4 meV/atom	ASE [2]
CASTEP	8.0	plane waves	OTFG CASTEP 8.0	0.5 meV/atom	CASTEP [7]
ABINIT	7.7.3	plane waves	PAW JTH v0.2	0.6 meV/atom	F. Jollet and M. Torrent

K. Lejaeghere, V. Van Speybroeck, G. Van Oost, and S. Cottenier, Crit. Rev. Solid State Mater. Sci. 39, 1-24 (2014); https://molmod.ugent.be/deltacodesdft. Reference code: WIEN2k

Veracity – Approximate Treatment of Exchange-Correlation

Perdew's Dream: A Jacob's Ladder

The exchange-correlation functional

accuracy ↑

5	unoccupied $\psi_i(\mathbf{r})$,
4	occupied $\psi_i(\mathbf{r})$,
3	$\tau(\mathbf{r})$,
2	$\nabla n(\mathbf{r})$,
1	$n(\mathbf{r})$,

our favorite

EX + cRPA, as ...

hybrids (...

m... ...nation

...ation

Functionals of level 1 and 2 suffer from severe self-interaction errors.
Functionals of level 1, 2, 3, & 4 are lacking the long-range vdW tails.
With "Level 5 plus" validation (error estimation) is becoming possible.

$\tau(\mathbf{r})$: Kohn-Sham ...

EX: exact ... $\frac{\psi_m(\mathbf{r})\psi_m^*(\mathbf{r}')\psi_n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}$

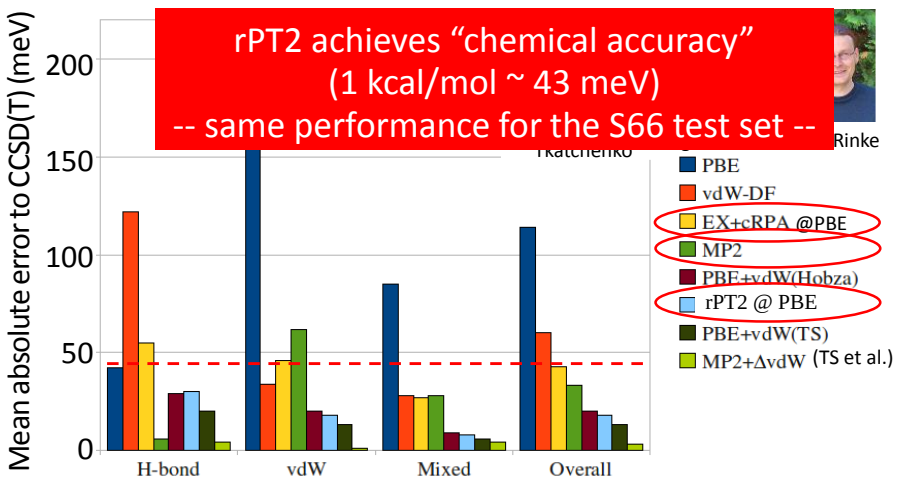
cRPA : ... correlation

ACFD : ad... connec... ...ation dissipation theorem

Bohm, Pines (1953); Gell-Mann, ...ueckner (1957);
Gunnarsson, Lundqvist (1975, 1976); Langreth, Perdew (1977);
X. Ren, P. Rinke, C. Joas, and M. S., Invited Review, Mater. Sci. 47, 21 (2012)

Performance of rPT2 for Weak Intermolecular Interactions:

S22 Test Set



CCSD(T): *Jurecka, Sponer, Cerny, Hobza,* ***PCCP*** *(2006).* ***Langreth-Lundqvist :*** *Gulans, Puska, Nieminen,* ***PRB*** *(2009);*
rPT2: *X. Ren et al.* ***PRL*** *(2011) and NJP (2013).* ***TS:*** *A. Tkatchenko and M.S.,* ***PRL*** *(2009); A. Tkatchenko et al.,* ***JCP*** *(2009)*

Test Sets for Materials Science and Engineering?

Chemists have shown the way. For small and light molecules they developed test sets: G2, NHTBH38, HTBH38, S22, S66 ...

We need a materials test set! We can now do renormalized second-order perturbation theory (similar to CCSD) and even full CI (*) – for certain systems.

Comparison with experiment is very important as well (adsorption energies of molecules, *e.g.* by microcalometry). However, theory-theory comparison is better defined.

(*) G. H. Booth, A. J. W. Thom, and A. Alavi, J. Chem. Phys. 131, 054106 (2009).
G. H. Booth, A. Grüneis, G. Kresse, and A. Alavi, Nature 493, 365 (2013).

Test set for materials science and engineering

MSE TEST SET

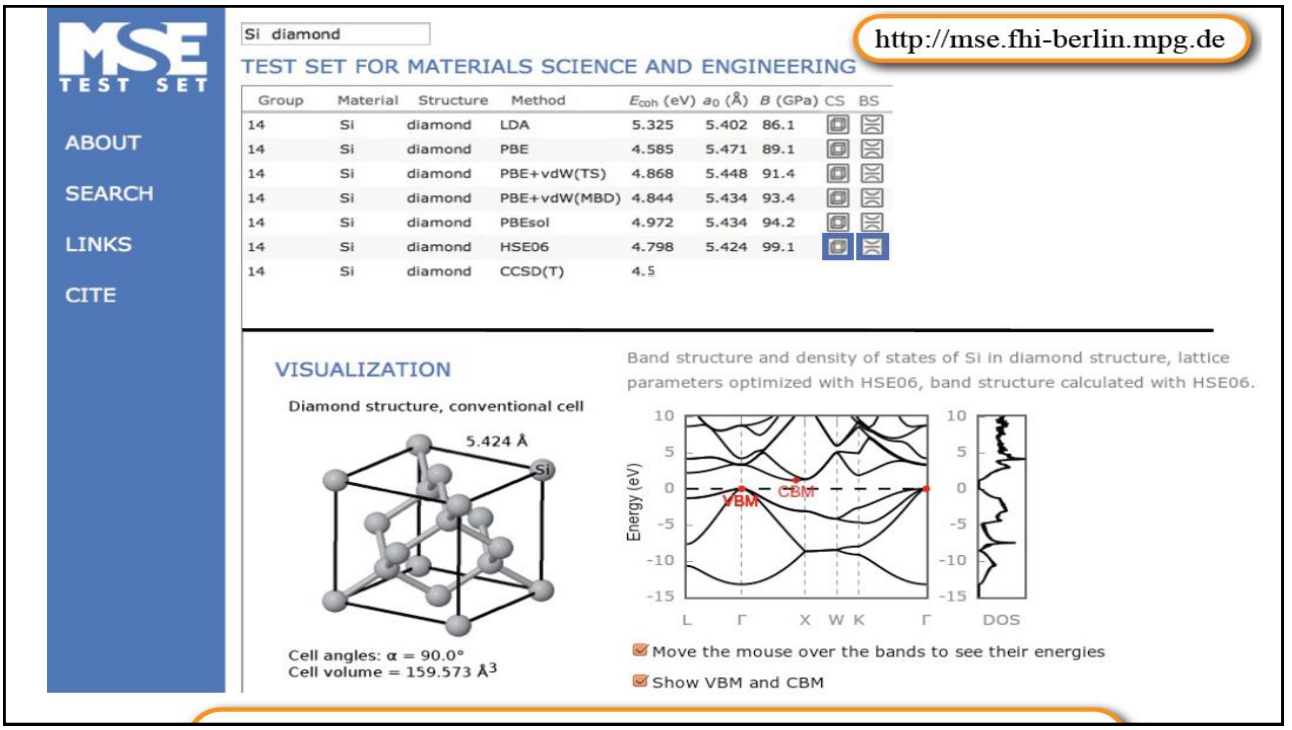
7 elements and 12 binaries with cubic structure (for the start)

Light main group elements

H						He
Li	Be	B	C	N	O	F	Ne
Na	Mg	Al	Si	P	S	Cl	Ar
K	Ca	Ga	Ge	As	Se	Br	Kr
Rb	Sr	In	Sn	Sb	Te	I	Xe
Cs	Ba						

Ne, Ar, Al (fcc); Li, Na (bcc); C, Si (diamond);
LiH, LiF, LiCl, NaF, NaCl, MgO, MgS (rocksalt);
BeS, BP, AlP, SiC, BN (zincblende)

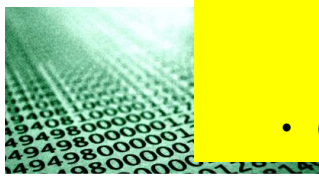
- **MSE properties**: cohesive, electronic, elastic and vibrational
- **Representative** for cubic metals, semiconductors, and insulators
- **Numerically accurate reference values from theory**, incl. MP2, RPA, CCSD(T)



The Four *V* of Big Data and an *A*

Data – data – data (analog to Moore's law)
numbers, arrays, figures, movies, …

Computed data: Query and read out what was stored. (high-throughput screening)
Shouldn't we do more?!



Big-Data Analytics

- Finding ***d*** from ***P***
- Causality: the science behind ***P***(***d***)

Big-Data Challenge: *"four V"*:

Volume (amount of data),

Variety (heterogeneity of form and meaning of data),

Veracity (uncertainty of data quality),

Velocity at which data may change or new data arrive.

The four V should be complemented by an "***A***", **the Big-Data Analytics**:

- identify (so far) hidden correlations,
- which materials should be studied next as most promising candidates,
- identify anomalies,
- identify the mechanisms that govern a certain material property or function.

Kernel Regression

We have data $\{\boldsymbol{P}_i\}$ at "coordinates" $\{\boldsymbol{x}_i\}$ $\quad \boldsymbol{x}_i$ = set of descriptive parameters (descriptor)

$$\boldsymbol{P}_i = \boldsymbol{P}(\boldsymbol{x}_i) = \sum_{k=1}^{N} c_k K(\boldsymbol{x}_i, \boldsymbol{x}_k)$$

Linear regression: $\quad K(\boldsymbol{x}_i, \boldsymbol{x}_k) = \boldsymbol{x}_i \,.\, \boldsymbol{x}_k \qquad \boldsymbol{P}(\boldsymbol{x}_i) = \boldsymbol{x}_i \,.\, \boldsymbol{c}^*$

Polynomial kernel $\quad K(\boldsymbol{x}_i, \boldsymbol{x}_k) = (\boldsymbol{x}_i \,.\, \boldsymbol{x}_k + c)^d$

Gaussian kernel $\quad K(\boldsymbol{x}_i, \boldsymbol{x}_k) = \exp\left(-\sum_j (\boldsymbol{x}_i - \boldsymbol{x}_k)^2 / 2\sigma_j^2\right)$

More data means better representation.
Do we "learn" anything?
For successful learning, we need a "good" descriptor: $\boldsymbol{P}(\boldsymbol{x}_i) \rightarrow \boldsymbol{P}(\boldsymbol{d}_i)$

Toy Model: Descriptor for the Classification "Zincblende/Wurtzite or Rocksalt?"

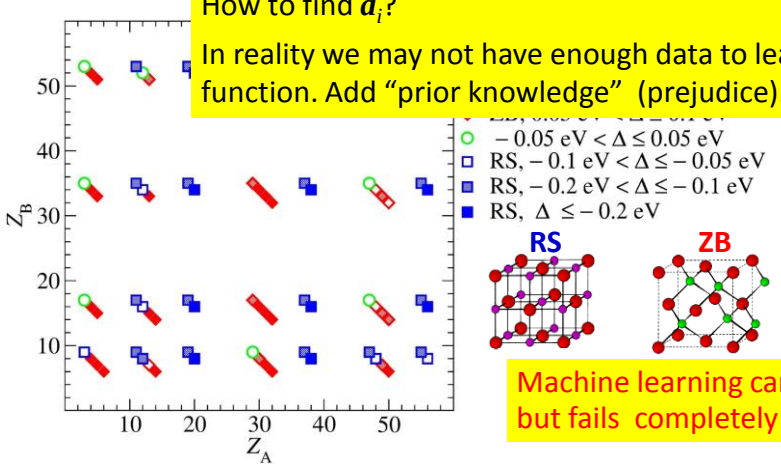
Only DFT- … Z_B?

82 oct…

Arrange data $\boldsymbol{P}(\boldsymbol{x}_i) \rightarrow \boldsymbol{P}(\boldsymbol{d}_i)$, so that $\boldsymbol{P}(\boldsymbol{d}_i)$ is a "well behaved function". Fit this function by machine learning.

How to find $\boldsymbol{d}_i$?

In reality we may not have enough data to learn a complex function. Add "prior knowledge" (prejudice).



…between … are very small. For Si: 0.01% of the energy of a Si atom, or 0.1% of the 4 valence electrons. Complexity: $T_s[n]$ and E_{xc}.

Machine learning can fit the $P(Z_A, Z_B)$ data well, but fails completely in predictions.

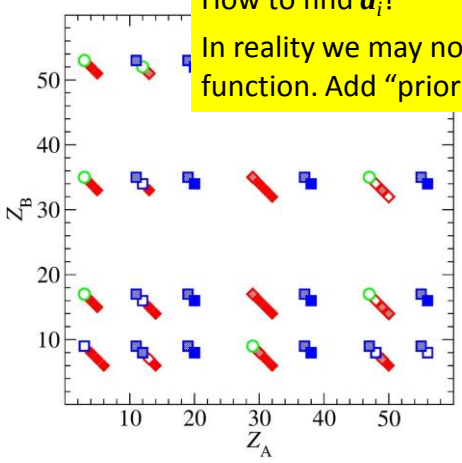
Toy Model: Descriptor for the Classification "Zincblende/Wurtzite or Rocksalt?"

Arrange data $\boldsymbol{P}(\boldsymbol{x}_i) \rightarrow \boldsymbol{P}(\boldsymbol{d}_i)$, so that $\boldsymbol{P}(\boldsymbol{d}_i)$ is a "well behaved function". Fit this function by machine learning.

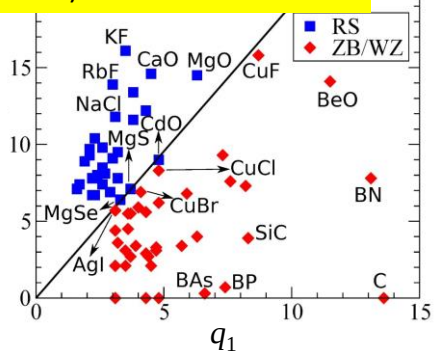
How to find $\boldsymbol{d}_i$?

In reality we may not have enough data to learn a complex function. Add "prior knowledge" (prejudice).

82 oct



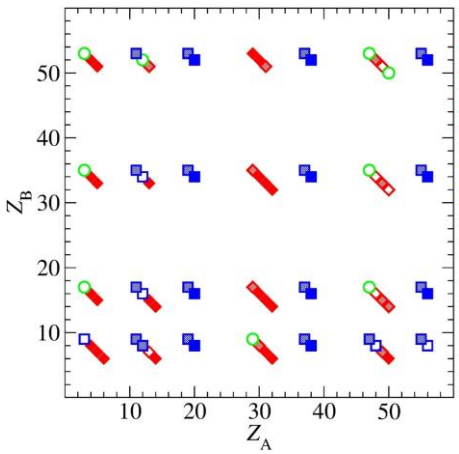
?



Toy Model: Descriptor for the Classification "Zincblende/Wurtzite or Rocksalt?"

Only DFT-LDA: Can we predict not yet calculated LDA structures from Z_A and Z_B?

82 octet AB binary compounds



Wish List for a Descriptor

a) ➕ A descriptor is an array of real numbers that uniquely characterizes the material as well as property-relevant elementary processes.

b) ➖ Materials that are very different (similar) should be characterized by very different (similar) descriptor values.

c) ➕ The determination of the descriptor must not involve calculations as intensive as those needed for the evaluation of the property to be predicted.

d) ➖ The dimension of the descriptor should be as low as possible, but not lower.

Toy Model: Descriptor for the Classification "Zincblende/Wurtzite or Rocksalt?"

the key scientific challenge: find a descriptor $\boldsymbol{d}$ that works.

How to find a good descriptor, also for more complicated properties and functions?

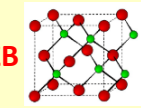
Example (the traditional way):

J. C. Phillips and J. A. Van Vechten (1969/70)(*) :
A two-dimensional descriptor that distinguishes materials that crystallize in ZB/WZ vs. RS structures:

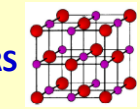
$E_h =$
$C =$

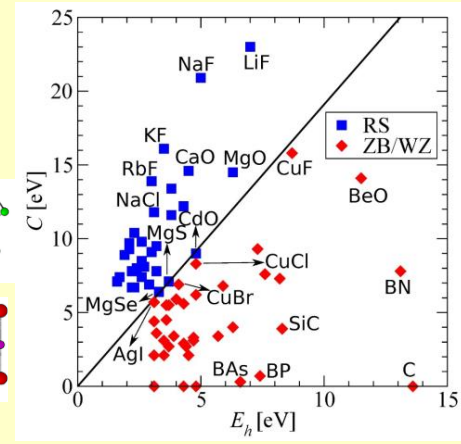
related to crystal's band gap, dielectric constant, nearest-neighbor distance

ZB



RS





There are several other descriptors for this classification goal, by various authors.

(*) J. A. Van Vechten, Phys. Rev. B **182** , 891 (1969);
J. C. Phillips, Rev. Mod. Phys. **42**, 317 (1970)

Statistical Learning (Machine Learning)

fit and/or interpolation of discrete, known data points $\{ P_i \}$ and building a function $P(\boldsymbol{d})$

the key scientific challenge: find a reliable, low dimensional descriptor $\boldsymbol{d}$.

kernel ridge regression		linear
		R. Tibshirani, J. Royal Statist. Soc. B 58, 267 (1996)
$P(\boldsymbol{d}) = \sum_{i=1}^{N} c_i \exp\left(-\lVert \boldsymbol{d}_i - \boldsymbol{d} \rVert_2^2 / 2\sigma^2\right)$		$P(\boldsymbol{d}) = \boldsymbol{d}\boldsymbol{c}$
$\sum_{i=1}^{N} (P(\boldsymbol{d}_i) - P_i)^2 \quad +$	minimize	$\sum_{i=1}^{N} (P(\boldsymbol{d}_i) - P_i)^2 \quad +$
$\lambda \sum_{i,j=1}^{N,N} c_i c_j \exp\left(-\lVert \boldsymbol{d}_i - \boldsymbol{d}_j \rVert_2^2 / 2\sigma^2\right)$		$\lambda \lVert \boldsymbol{c} \rVert_1$
$\lVert \boldsymbol{d}_i - \boldsymbol{d}_j \rVert_2^2 = \sum_{\alpha=1}^{\Omega} (d_{i,\alpha} - d_{j,\alpha})^2$		$\lVert \boldsymbol{c} \rVert_1 = \sum_{\alpha=1}^{M} \lvert c_\alpha \rvert$
		least absolute shrinkage and selection operator (LASSO) for feature selection

1) Primary Features, 2) Feature Space, 2) Descriptors

1)

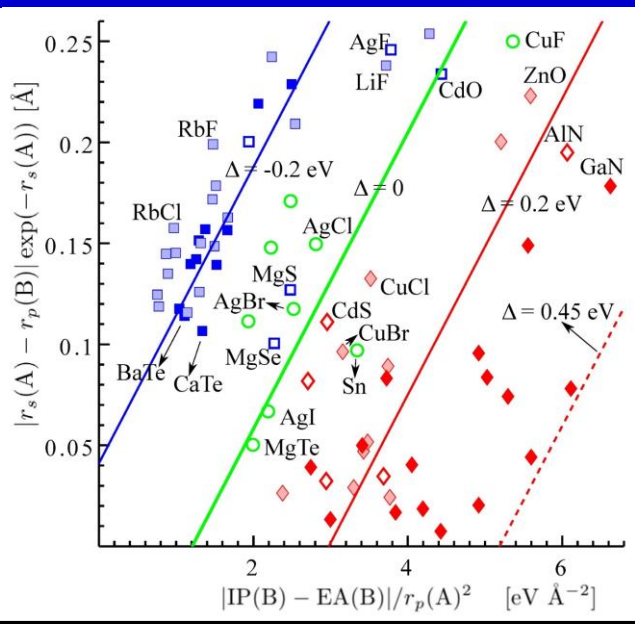
ID	Description **free atoms**	Symbols	#
$A1$	Ionization Potential (IP) and Electron Affinity (EA)	IP(A) EA(A) IP(B) EA(B) [1]	4
$A2$	Highest occupied (H) and lowest unoccupied (L) Kohn-Sham levels	H(A) L(A) H(B) L(B)	4
$A3$	Radius at the max. value of s, p, and d valence radial radial probability density	r_s(A) r_p(A) r_d(A) r_s(B) r_p(B) r_d(B)	6

ID	Description **free dimers**	Symbols	#
$A4$	Binding energy	E_b(AA) E_b(BB) E_b(AB)	3
$A5$	HOMO-LUMO KS gap	HL(AA) HL(BB) HL(AB)	3
$A6$	Equilibrium distance	d(AA) d(BB) d(AB)	3

2) We start with 23 primary features
and build > 10,000 non linear combinations

3) LASSO finds the most important descriptors: $\frac{\mathrm{IP(B)}-\mathrm{EA(B)}}{r_p(\mathrm{A})^2}$, $\frac{|r_s(\mathrm{A})-r_p(\mathrm{B})|}{\exp(r_s(\mathrm{A}))}$, $\frac{|r_p(\mathrm{B})-r_s(\mathrm{B})|}{\exp(r_d(\mathrm{A})+r_s(\mathrm{B}))}$

Statistical Learning (Machine Learning): LASSO, 2-Dim. Descriptor



$\Delta = E(\mathrm{RS}) - E(\mathrm{ZB})$
- ZB, $\Delta > 0.2$ eV
- ZB, 0.1 eV $< \Delta \leq 0.2$ eV
- ZB, 0.05 eV $< \Delta \leq 0.1$ eV
- -0.05 eV $< \Delta \leq 0.05$ eV
- RS, -0.1 eV $< \Delta \leq -0.05$ eV
- RS, -0.2 eV $< \Delta \leq -0.1$ eV
- RS, $\Delta \leq -0.2$ eV

$$P(\boldsymbol{d}) = \boldsymbol{dc}$$

The complexity and science is in the descriptor (identified from >10,000 features).

Statistical Learning (Machine Learning): Descriptor

Mean absolute error (MAE), and maximum absolute error (MaxAE), in eV, (first two lines) and for a leave-10%-out cross validation (CV), averaged over 150 random selections of the training set (last two lines). For (Z_A*, Z_B*), each atom is identified by a string of three random numbers.

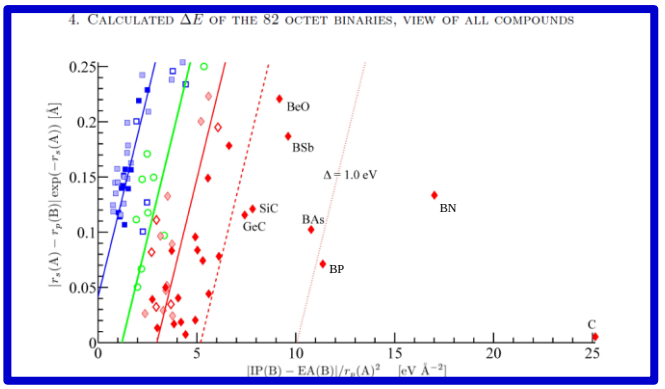
Descriptor	Z_A, Z_B	Z_A*, Z_B*	1D	2D	3D	5D
MAE	$1*10^{-4}$	$3*10^{-3}$	0.12	0.08	0.07	0.05
MaxAE	$8*10^{-4}$	0.03	0.32	0.32	0.24	0.20
MAE, CV	0.13	0.14	0.12	0.09	0.07	0.05
MaxAE, CV	0.43	0.42	0.27	0.18	0.16	0.12

Statistical Learning (Machine Learning): Descriptor

Mean absolute error (MAE), and maximum absolute error (MaxAE), in eV, (first two lines) and for a leave-10%-out cross validation (CV), averaged over 150 random selections of the training set (last two lines). For (Z_A*, Z_B*), each atom is identified by a string of three random numbers.

Descriptor	Z_A, Z_B	Z_A*, Z_B*	1D	2D	3D	5D
MAE	$1*10^{-4}$	$3*10^{-3}$	0.12	0.08	0.07	0.05
MaxAE	$8*10^{-4}$	0.03	0.32	0.32	0.24	0.20
MAE, CV	0.13	0.14	0.12	0.09	0.07	0.05
MaxAE, CV	0.43	0.42	0.27	0.18	0.16	0.12

Statistical Learning (Machine Learning): Descriptor



lute error (MaxAE), in eV, (first two
CV), averaged over 150 random
(Z_A*, Z_B*), each atom is identified by

			1D	2D	3D	5D
			0.12	0.08	0.07	0.05
			0.32	0.32	0.24	0.20
MAE, CV	0.13	0.14	0.12	0.09	0.07	0.05
MaxAE, CV	0.43	0.42	0.27	0.18	0.16	0.12

Drawing Causal Inference from Big Data (Scientific Insight)

Correlation between ***d*** and ***P*** , i.e. ***P*** is a function of ***d***, ***P***(***d***), reflects causal inference

if it is based on sufficient information(*)

There are four possibilities (types of causality) behind ***P***(***d***):


Judea Pearl

1. ***d*** → ***P*** : ***P*** "listens" to ***d***
2. ***A*** → ***d*** and ***A*** → ***P*** : There is no direct connection between ***d*** and ***P***, but ***d*** and ***P*** both "listen" to a third "actuator"
3. ***P*** → ***d*** : ***d*** "listens" to ***P***
4. There is no direct connection between ***d*** and ***P***, but they have a common effect that listens to both and screams: "I occurred" (Berkson bias; Judea Pearl)

(*) Construct ***d*** with scientific knowledge (prejudice?), or use "big data" for {P_i }.

Drawing Causal Inference from Big Data (Scientific Insight)

Example:

The probability of childhood leukemia is higher for people living close to electricity power lines.

There is no direct connection between leukemia and the electromagnetic field.

Living close to electric power lines is not a desired residence. People living near power lines tend to be poorer than the control group, and there is a relationship between poverty and cancer.

Poverty → higher probability for living close to power lines

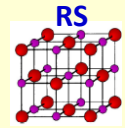
Poverty → higher chances for cancer

? ? correlation
no direct relation
causality

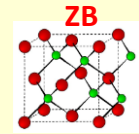
Drawing Causal Inference from Big Data (Scientific Insight)

Our previous example:
Prediction of the energy difference between ZB/WZ and RS of binary Compound semiconductors

RS



ZB



There is no scientific law that connects the descriptor

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

directly with the total-energy difference (we are not able to write it down).
However, Z_A, Z_B determine these descriptors,
and Z_A, Z_B determine the many-body Hamiltonians and the total-energy difference.

Poverty → higher probability for living close to power lines

Poverty → higher chances for cancer

? ? correlation
no direct relation
causality

Drawing Causal Inference from Big Data (Scientific Insight)

Correlation between $\boldsymbol{d}$ and $\boldsymbol{P}$ reflects causal inference
if it is based on sufficient information(*)


Judea Pearl

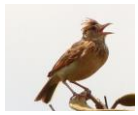
There are four possibilities (types of causality) behind $\boldsymbol{P}(\boldsymbol{d})$:

1. $\boldsymbol{d} \rightarrow \boldsymbol{P}$: $\boldsymbol{P}$ "listens" to $\boldsymbol{d}$
2. $\boldsymbol{A} \rightarrow \boldsymbol{d}$ and $\boldsymbol{A} \rightarrow \boldsymbol{P}$: There is no direct connection between $\boldsymbol{d}$ and $\boldsymbol{P}$, but $\boldsymbol{d}$ and $\boldsymbol{P}$ both "listen" to a third "actuator"
3. $\boldsymbol{P} \rightarrow \boldsymbol{d}$: $\boldsymbol{d}$ "listens" to $\boldsymbol{P}$
4. There is no direct connection between $\boldsymbol{d}$ and $\boldsymbol{P}$, but they have a common effect that listens to both and screams: "I occurred" (Berkson bias; Judea Pearl)

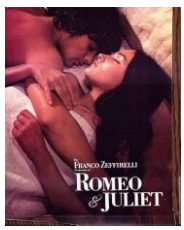
(*) Construct $\boldsymbol{d}$ with scientific knowledge (prejudice?), or use "big data" for $\{\boldsymbol{P}_i\}$.

Drawing Causal Inference from Big Data (Scientific Insight)

case # 3



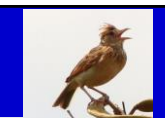
ROMEO: "It was the lark, the bird that sings at dawn, not the nightingale. Look, my love, what are those streaks of light in the clouds parting in the east? Night is over, and day is coming. … "



The ***singing of the lark*** is a good descriptor for "***the sun will rise soon***".
The ***singing of the lark*** is not the actuator of (the mechanism behind) the sunrise.

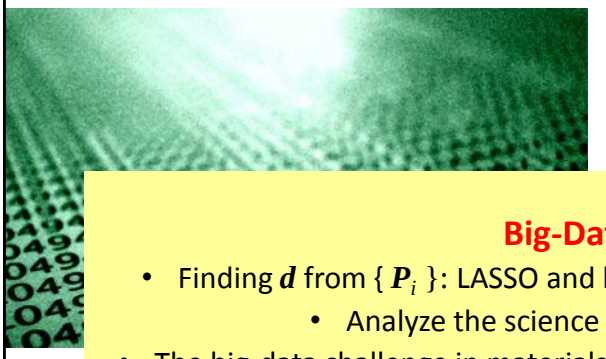
Conclusion / Suggestion: Accept "larks" (not just scientific laws) to predict materials properties.

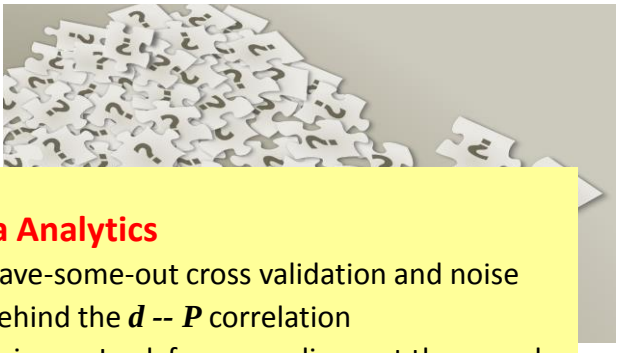
The Four *V* of Big Data and an *A*



Data – data – data (analog to Moore's law)
numbers, arrays, figures, movies, …

Computed data: Query and read out what was stored. (high-throughput screening)
Shouldn't we do more?!





Big-Data Analytics

- Finding $\boldsymbol{d}$ from $\{ \boldsymbol{P}_i \}$: LASSO and leave-some-out cross validation and noise
- Analyze the science behind the $\boldsymbol{d}$ -- $\boldsymbol{P}$ correlation
- The big-data challenge in materials science: Look for anomalies, not the crowd

Summary and Outlook

- Machine learning may find structure in data that is invisible to humans.
- Causal models, i.e. finding causal descriptors, are richer. They are able to provide scientific insight and understanding.
- They can tell how to do machine learning on difficult tasks.
- Question: Why do we want to achieve insight and understanding? Isn't it good enough to have a predictive model?
- Question: Is it possible to assign error bars to predictions (of unexpected situations)?

Next steps:

- Beyond the linear fit: Non-linear kernel with l_1-norm regularization.
- Higher accuracy: MaxAE loss function with l_1-norm regularization.
- Improving the systematic creation of the feature space.