NOMAD

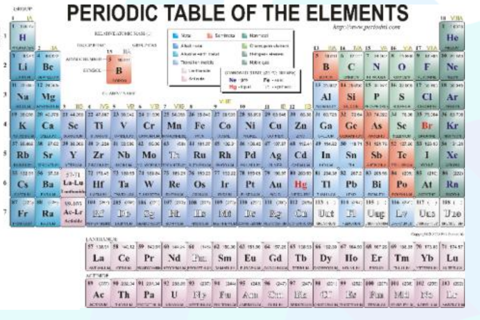
NOVEL MATERIALS DISCOVERY

Patterns, Correlations, and Causality in Big Data of Materials: Analytics for Novel Materials Discovery

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

Dmitri Mendeleev (1834-1907)

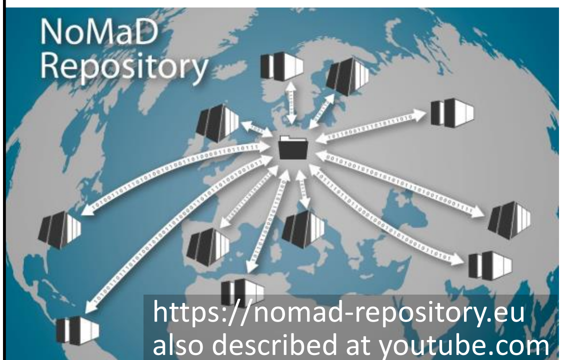
- crystal-structure prediction
- figure of merit of thermoelectrics (as function of *T*)
- turn-over frequency of catalytic materials (as function of *T* and *p*)
- efficiency of photovoltaic systems
- etc.



(*) Work performed in collaboration with
Luca Ghiringhelli, Jan Vybiral, Claudia Draxl, et al.

NOMAD
NOVEL MATERIALS DISCOVERY

The NOMAD Laboratory
A European Centre of Excellence



The NoMaD Repository accepts (and requests) in- and output files of all important codes. Currently, the NoMaD Repository contains **3,026,745** entries.

http://NOMAD-CoE.eu

---

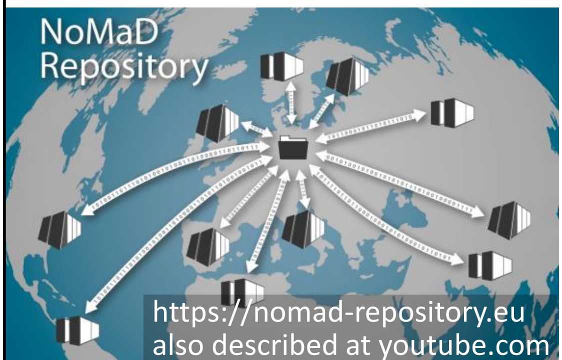
https://www.youtube.com/watch?v=L-nmRSH4NQM
http://v.youku.com/v_show/id_XMTM0NDA0NDIxMg==.html

NoMaD

The Novel Materials Discovery Repository

The NOMAD Laboratory
A European Centre of Excellence

NOMAD
NOVEL MATERIALS DISCOVERY



The NoMaD Repository accepts (and requests) in- and output files of all important codes. Currently, the NoMaD Repository contains **3,026,745** entries.

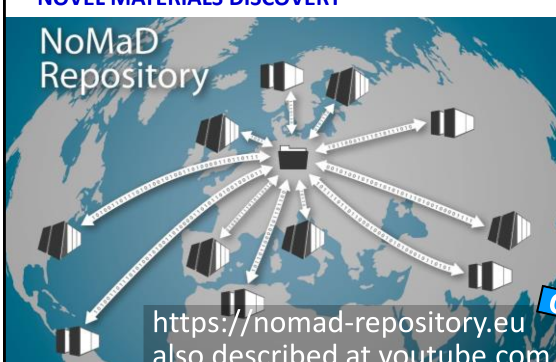
There are 30-40 important codes used in computational materials science.

Nomenclature, data representation, and file formats of the input and output files of these codes are different. The heterogeneity could hardly be worse.

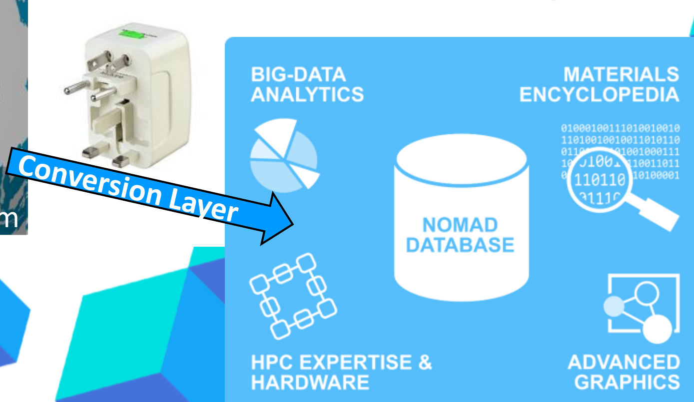
http://NOMAD-CoE.eu

The NOMAD Laboratory
A European Centre of Excellence

NOMAD
NOVEL MATERIALS DISCOVERY



The NoMaD Repository accepts (and requests) in- and output files of all important codes. Currently, the NoMaD Repository contains **3,026,745** entries.

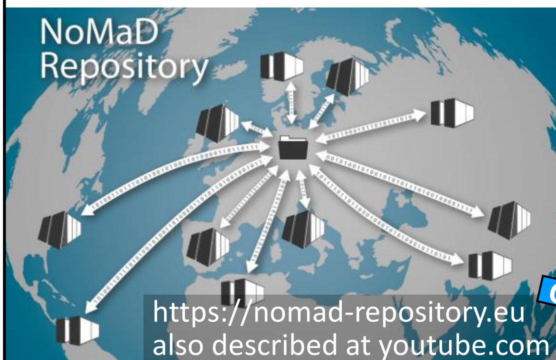


http://NOMAD-CoE.eu

NOMAD

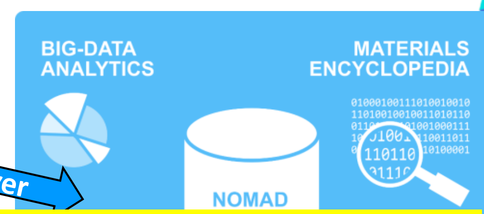
NOVEL MATERIALS DISCOVERY

The NOMAD Laboratory
A European Centre of Excellence



The NoMaD Repository accepts (and requests) in- and output files of all important codes. Currently, the NoMaD Repository contains **3,026,745** entries.

Conversion Layer



Why do we consider "all" electronic structure theory codes in our NOMAD repository and data base?

There are millions of data already. **Do science with the data! Select the studied systems more intelligently.**

NOMAD

NOVEL MATERIALS DISCOVERY

The Big-Data Challenge
The Amount of Data Created Every Day Is Significant

Most Data Are Not Used And Even Thrown Away

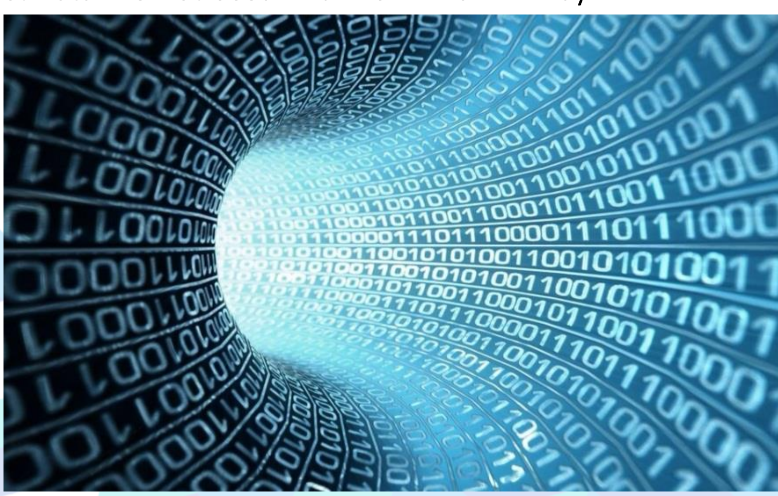
The Four V and an A !

Volume (amount of data),

Variety (heterogeneity of form and meaning of data),

Velocity at which data may change or new data arrive,

Veracity (uncertainty of quality).



NOMAD
NOVEL MATERIALS DISCOVERY

The Big-Data Challenge

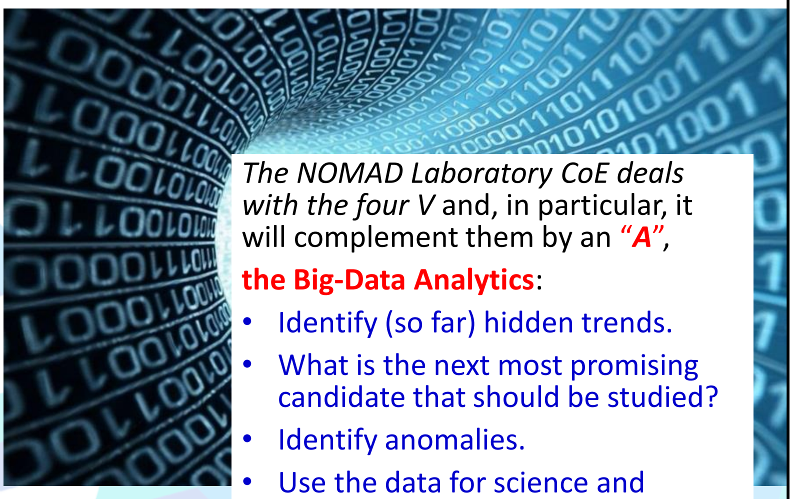
The Amount of Data Created Every Day Is Significant

Most Data Are Not Used And Even Thrown Away

The Four V and an A !

Volume (amount of data),
Variety (heterogeneity of form and meaning of data),
Velocity at which data may change or new data arrive,
Veracity (uncertainty of quality).

Query and read out what was stored;
high-throughput screening.



The NOMAD Laboratory CoE deals with the four V and, in particular, it will complement them by an *"A"*,
the Big-Data Analytics:

- Identify (so far) hidden trends.
- What is the next most promising candidate that should be studied?
- Identify anomalies.
- Use the data for science and engineering, for R & D.

NOMAD
NOVEL MATERIALS DISCOVERY

Science **351**, aad3000 (2016)

RESEARCH

RESEARCH ARTICLE

DFT METHODS

Reproducibility in density functional theory calculations of solids

Kurt Lejaeghere,[1*] Gustav Bihlmayer,[2] Torbjörn Björkman,[3,4] Peter Blaha,[5]

Science **351**, aad3000 (2016)

Kurt Lejaeghere,[1*] Gustav Bihlmayer,[2] Torbjörn Björkman,[3,4] Peter Blaha,[5] Stefan Blügel,[2] Volker Blum,[6] Damien Caliste,[7,8] Ivano E. Castelli,[9] Stewart J. Clark,[10] Andrea Dal Corso,[11] Stefano de Gironcoli,[11] Thierry Deutsch,[7,8] John Kay Dewhurst,[12] Igor Di Marco,[13] Claudia Draxl,[14,15] Marcin Dułak,[16] Olle Eriksson,[13] José A. Flores-Livas,[12] Kevin F. Garrity,[17] Luigi Genovese,[7,8] Paolo Giannozzi,[18] Matteo Giantomassi,[19] Stefan Goedecker,[20] Xavier Gonze,[19] Oscar Grånäs,[13,21] E. K. U. Gross,[12] Andris Gulans,[14,15] François Gygi,[22] D. R. Hamann,[23,24] Phil J. Hasnip,[25] N. A. W. Holzwarth,[26] Diana Iuşan,[13] Dominik B. Jochym,[27] François Jollet,[28] Daniel Jones,[29] Georg Kresse,[30] Klaus Koepernik,[31,32] Emine Küçükbenli,[9,11] Yaroslav O. Kvashnin,[13] Inka L. M. Locht,[13,33] Sven Lubeck,[14] Martijn Marsman,[30] Nicola Marzari,[9] Ulrike Nitzsche,[31] Lars Nordström,[13] Taisuke Ozaki,[34] Lorenzo Paulatto,[35] Chris J. Pickard,[36] Ward Poelmans,[1,37] Matt I. J. Probert,[25] Keith Refson,[38,39] Manuel Richter,[31,32] Gian-Marco Rignanese,[19] Santanu Saha,[20] Matthias Scheffler,[15,40] Martin Schlipf,[22] Karlheinz Schwarz,[5] Sangeeta Sharma,[12] Francesca Tavazza,[17] Patrik Thunström,[41] Alexandre Tkatchenko,[15,42] Marc Torrent,[28] David Vanderbilt,[23] Michiel J. van Setten,[19] Veronique Van Speybroeck,[1] John M. Wills,[43] Jonathan R. Yates,[29] Guo-Xu Zhang,[44] Stefaan Cottenier[1,45*]

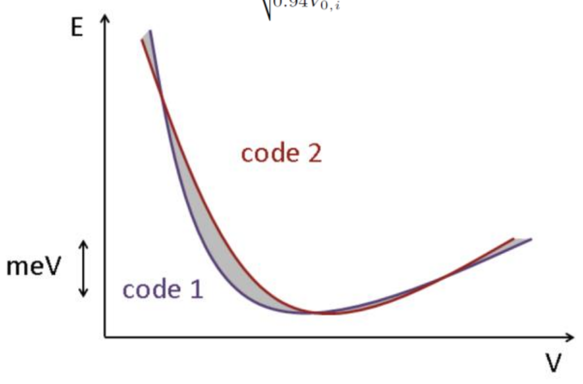
Science **351**, aad3000 (2016)

NOVEL MATERIALS DISCOVERY

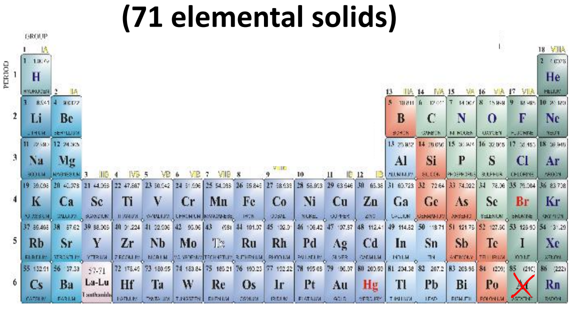
Δ-Value Project in Materials Science

Science 351, aad3000 (2016)

$$\Delta_i(a,b) = \sqrt{\int_{0.94V_{0,i}}^{1.06V_{0,i}} \frac{\left(E_{b,i}(V) - E_{a,i}(V)\right)^2}{0.12V_{0,i}} \, \mathrm{d}V}$$



An exhaustive test set (71 elemental solids)



NOVEL MATERIALS DISCOVERY

Comparing Solid State DFT Codes, Basis Sets, and Potentials (K. Lejaeghere et al.)

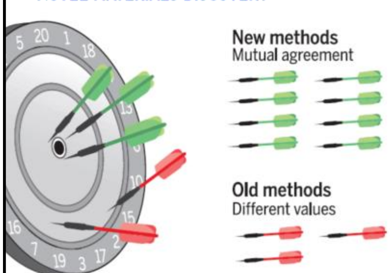
https://molmod.ugent.be/deltacodesdft
Science 351, aad3000 (2016)

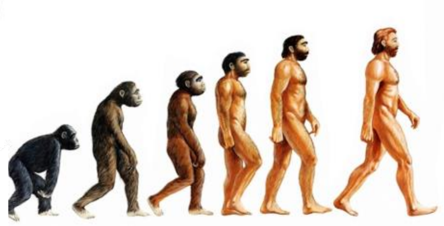
Code	Version	Basis	Electron treatment	Δ-value meV/atom
WIEN2k	13.1	LAPW/APW+lo	all-electron	0
FHI-aims	081213	tier2 numerical orbitals	all-electron (relativistic atomic_zora scalar)	0.2
Exciting	development version	LAPW+xlo	all-electron	0.2
Quantum ESPRESSO	5.1	plane waves	SSSP Accuracy (mixed NC/US/PAW potential lib.)	0.3
Elk	3.1.5	APW+lo	all-electron	0.3
VASP	5.2.12	plane waves	PAW 2015	0.4
FHI-aims	081213	tier2 numerical orbitals	all-electron (relativistic zora scalar 1e-12)	0.4
CASTEP	9.0	plane waves	OTFG CASTEP 9.0	0.5

NOVEL MATERIALS DISCOVERY

Comparing Solid State DFT Codes, Basis Sets, and Potentials (K. Lejaeghere et al.)

Science 351, aad3000 (2016)



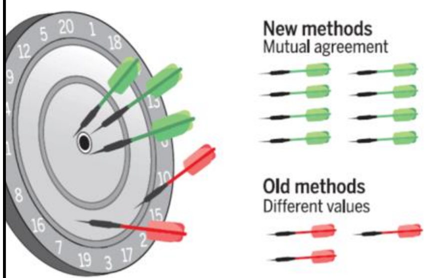


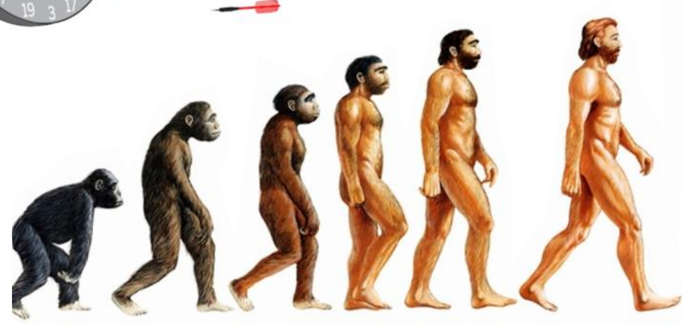
O_2-crystal

	exciting	FHI-AIMS	FLEUR	FPLO	RSPt	WIEN2k	VASP201	VASP2012/VASP
exciting		0.1	0.4	1.8	0.7	0.6	7.7	0.2
FHI-AIMS/tier2	0.1		0.3	1.7	0.6	0.5	7.6	0.1
FLEUR	0.4	0.3		1.4	0.3	0.3	7.2	0.2
FPLO/T+F+s	1.8	1.7	1.4		1.1	1.2	5.8	1.6
RSPt	0.7	0.6	0.3	1.1		0.1	6.9	0.5
WIEN2k/acc	0.6	0.5	0.3	1.2	0.1		7.1	0.4
VASP2011/VASP	7.7	7.6	7.2	5.8	6.9	7.1		7.5
VASP2012/VASP	0.2	0.1	0.2	1.6	0.5	0.4	7.5	

Comparing Solid State DFT Codes, Basis Sets, and Potentials (K. Lejaeghere et al.)

Science 351, aad3000 (2016)





	year	⟨Δ⟩ vs AE
JTH01/ABINIT	2013	1.1
JTH02/ABINIT	2014	0.6
Vdb/CASTEP	1998	6.5
OTFG7/CASTEP	2013	2.6
OTFG9/CASTEP	2015	0.7
GPAW06/GPAW	2010	3.6
GPAW09/GPAW	2012	1.6
PSlib031/QE	2013	1.7
PSlib100/QE	2013	1.0
VASP2007/VASP	2007	2.0
VASP2012/VASP	2012	0.8
VASPGW2015/VASP	2015	0.6

Comparing Solid State DFT Codes, Basis Sets, and Potentials (K. Lejaeghere et al.)

NOVEL MATERIALS DISCOVERY

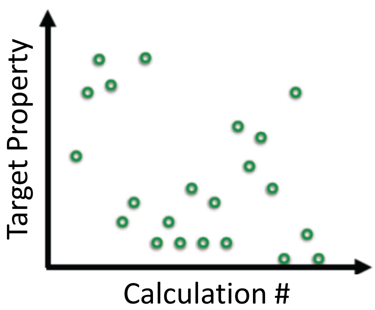
Use a code and settings (pseudopotentials (if necessary), basis set, etc.) that survived the Delta test.

Code	Version	Basis	Electron treatment	Δ-value meV/atom
WIEN2k	13.1	LAPW/APW+lo	all-elec	0
FHI-aims	081213	tier2 numerical orbitals		0.2
Exciting	development version	LAPW+xlo		0.2
Quantum ESPRESSO	5.1		Accuracy (mixed NC/US/PAW potential lib.)	0.3
Elk	3.1		all-electron	0.3
VASP	5.	ves	PAW 2015	0.4
FHI-aims	08	r2 numerical orbitals	all-electron (relativistic zora scalar 1e-12)	0.4
CASTEP	9.0	plane waves	OTFG CASTEP 9.0	0.5

A Delta value of < 0.5 is considered "perfect".
Most codes/pseudopotentials (now) perform with Δ-values between 1 and 2. Δ can be as high as 10.

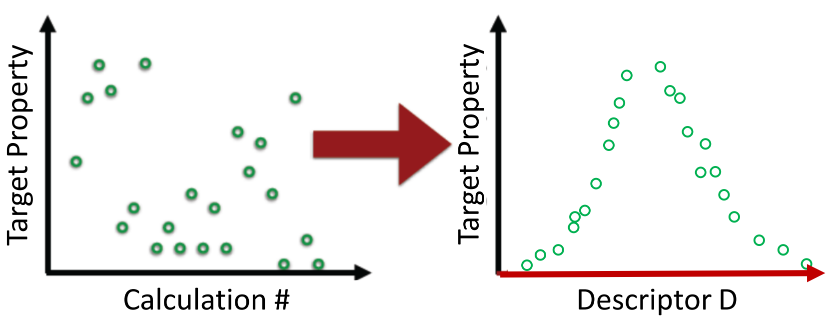
Find Structure in Big Data That Is *A Priori* “Not Visible”
Data Fitting, Statistical Learning, Machine Learning

Arrange/organize materials with respect to a property and a *set of simple descriptive parameters (a descriptor).*



Find Structure in Big Data That Is *A Priori* “Not Visible”
Data Fitting, Statistical Learning, Machine Learning

Arrange/organize materials with respect to a property and a *set of simple descriptive parameters (a descriptor).*

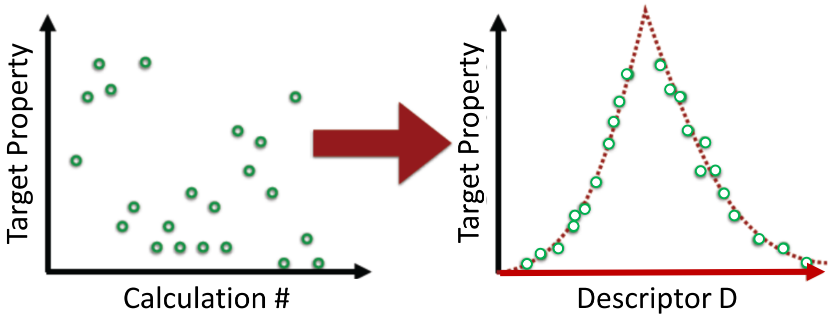


The descriptor can be designed: Rupp, von Lilienfeld, Behler, Csanyi, Seko, Tsuda, …

The descriptor can be selected out of a large set of candidates: Ozolins, Ghiringhelli.

Find Structure in Big Data That Is *A Priori* "Not Visible"
Data Fitting, Statistical Learning, Machine Learning

Arrange/organize materials with respect to a property and a *set of simple descriptive parameters (a descriptor).*



The descriptor can be designed: Rupp, von Lilienfeld, Behler, Csanyi, Seko, Tsuda, …

The descriptor can be selected out of a large set of candidates: Ozolins, Ghiringhelli.

More data means a better representation. Will we ever have enough data?

Big-Data-Driven Science vs. Model-Driven Science

NOVEL MATERIALS DISCOVERY

Traditional approach in the empirical sciences (e.g. physics, chemistry):

- Study a few systems
- Build a model,
- Improve the model when needed

(e.g. strength of transition metals Ti, … Fe, … Cu: d-band occupation, etc.).

The new option offered by Big-Data Analytics (and big-data-driven science):

- Find structure in big data that is probably invisible by standard tools.
- Offer many (thousands) of optional models and
- employ applied mathematics/information theory to find out which model is best (e.g. compressed sensing, statistical learning).

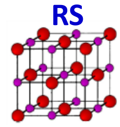
NOVEL MATERIALS DISCOVERY

Proof of Concept: Descriptor for the Classification “Zincblende/Wurtzite or Rocksalt?”

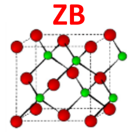
Crystal-structure prediction was and is one of the most important, basic challenges in materials science and engineering.

Can we predict not yet calculated structures from Z_A and Z_B? **Can we create a map: “The *ZB/W* community lives here and the *RS* community there?”**

RS



ZB



Energy differences between different structures are very small.

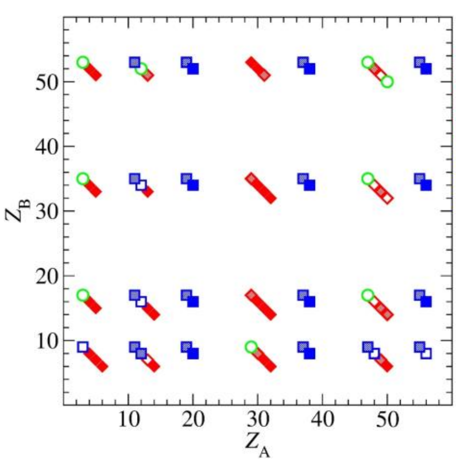
For Si: 0.01% of the energy of a Si atom, or 0.1% of the 4 valence electrons.

Complexity: $T_s[n]$ and E_{xc}.

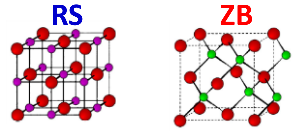
NOVEL MATERIALS DISCOVERY

Proof of Concept: Descriptor for the Classification “Zincblende/Wurtzite or Rocksalt?”

Crystal-structure prediction was and is one of the most important, basic challenges in materials science and engineering.



RS **ZB**



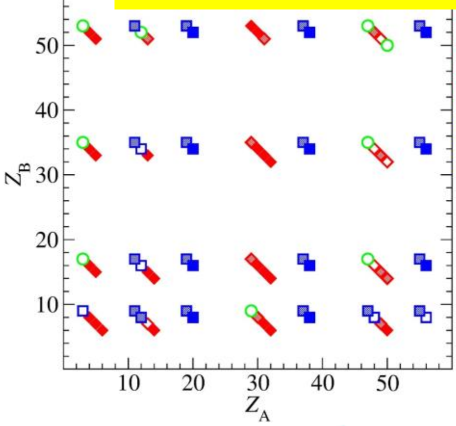
$\Delta = E(\mathrm{RS}) - E(\mathrm{ZB})$

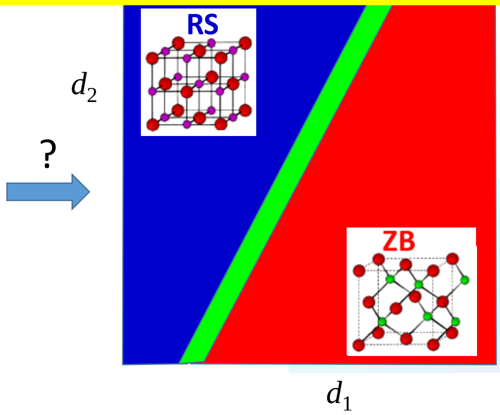
- ZB, $\Delta > 0.2$ eV
- ZB, 0.1 eV $< \Delta \le 0.2$ eV
- ZB, 0.05 eV $< \Delta \le 0.1$ eV
- -0.05 eV $< \Delta \le 0.05$ eV
- RS, -0.1 eV $< \Delta \le -0.05$ eV
- RS, -0.2 eV $< \Delta \le -0.1$ eV
- RS, $\Delta \le -0.2$ eV

Proof of Concept: Descriptor for the Classification "Zincblende/Wurtzite or Rocksalt?"

We need to arrange the data such that statistical learning is efficient. We need a good set of descriptive parameters.

How to find d_1, d_2?
In reality the representation will be higher than 2-dimensional.

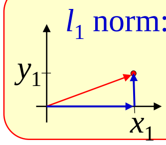




J. A. van Vechten, Phys. Rev. 182, 891 (1969).
J. C. Phillips, Rev. Mod. Phys. 42, 317 (1970).
A. Zunger, Phys. Rev. B 22, 5839 (1980).
D. G. Pettifor, Solid State Commun. 51, 31 (1984).
Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, Phys. Rev. B 85, 104104 (2012).

Statistical Learning (Machine Learning), Compressed Sensing

NOVEL MATERIALS DISCOVERY



l_1 norm: $|x_1| + |y_1|$ Manhattan (taxi cab) distance

l_2 norm: sqrt($x_1^2 + y_1^2$)

kernel ridge regression

$$P(\boldsymbol{d}) = \sum_{i=1}^{N} c_i \exp\left(-\|\boldsymbol{d}_i - \boldsymbol{d}\|_2^2/2\sigma^2\right)$$

minimize

$$\sum_{i=1}^{N}(P(\boldsymbol{d}_i) - P_i)^2 \; + \; \lambda \sum_{i,j=1}^{N,N} c_i c_j \exp\left(-\|\boldsymbol{d}_i - \boldsymbol{d}_j\|_2^2/2\sigma^2\right)$$

$$\|\boldsymbol{d}_i - \boldsymbol{d}_j\|_2^2 = \sum_{\alpha=1}^{\Omega}(d_{i,\alpha} - d_{j,\alpha})^2$$

linear

R. Tibshirani, J. Royal Statist. Soc. B 58, 267 (1996)

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

$$\sum_{i=1}^{N}(P(\boldsymbol{d}_i) - P_i)^2 \; + \; \lambda\|\boldsymbol{c}\|_1$$

$$\|\boldsymbol{c}\|_1 = \sum_{\alpha=1}^{M}|c_\alpha|$$

least absolute shrinkage and selection operator (LASSO) for feature selection

NOMAD 1) Primary Features, 2) Feature Space, 3) 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

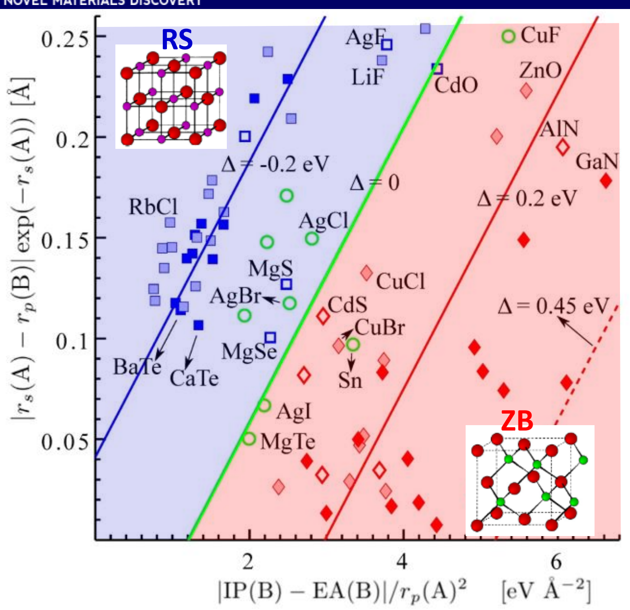
2) We start with 23 primary features and build > 10,000 non linear combinations

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

NOMAD "The Map" -- Compressed Sensing -- LASSO, 2-Dim. Descriptor

NOVEL MATERIALS DISCOVERY

L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, M.S., PRL ***114****, 105503 (2015).*



$\Delta = E(RS) - E(ZB)$
- ZB, $\Delta > 0.2$ eV
- ZB, 0.1 eV $< \Delta \le 0.2$ eV
- ZB, 0.05 eV $< \Delta \le 0.1$ eV
- -0.05 eV $< \Delta \le 0.05$ eV
- RS, -0.1 eV $< \Delta \le -0.05$ eV
- RS, -0.2 eV $< \Delta \le -0.1$ eV
- RS, $\Delta \le -0.2$ eV

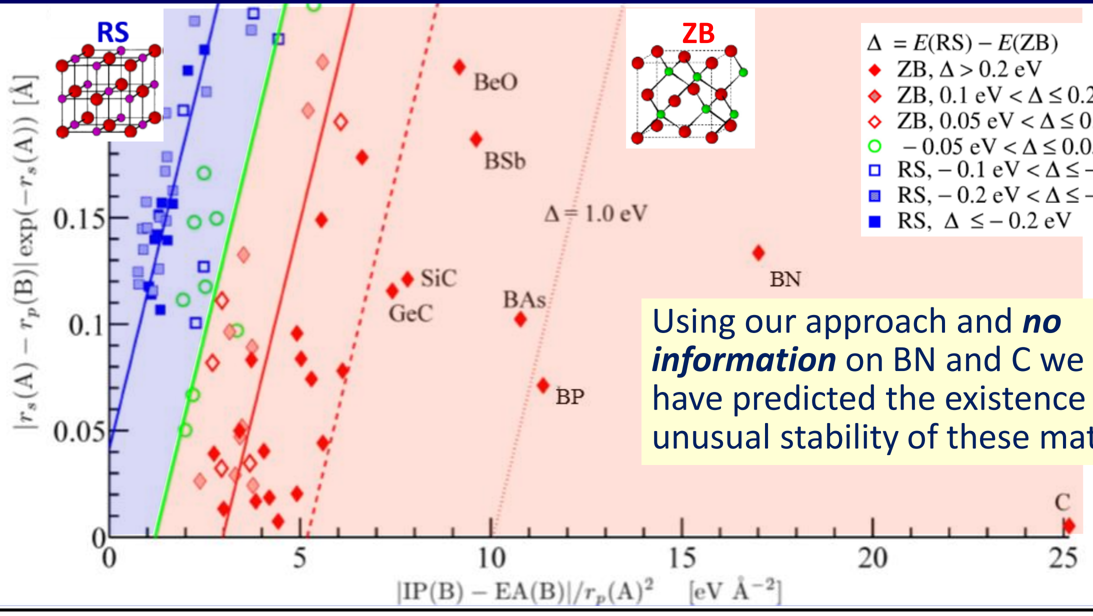
$$P(\mathbf{d}) = \mathbf{d}\mathbf{c}$$

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

NOVEL MATERIALS DISCOVERY

"The Map" -- Compressed Sensing -- LASSO, 2-Dim. Descriptor

L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, M.S., PRL ***114****, 105503 (2015).*



Using our approach and ***no information*** on BN and C we would have predicted the existence and unusual stability of these materials.

NOVEL MATERIALS DISCOVERY

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

NOMAD
NOVEL MATERIALS DISCOVERY

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

NOMAD
NOVEL MATERIALS DISCOVERY

Statistical Learning (Machine Learning): Descriptor

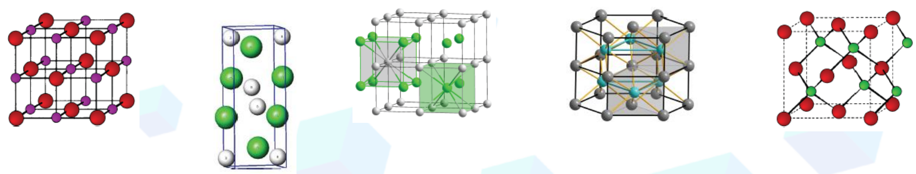
Mean absolute error (MAE), and maximum absolute error (MaxAE), in eV, (first two lines) and for a leave-10%-out cross validation (CV), averaged over 150 random selections of the training set (last two lines). For (Z_A*, Z_B*), each atom is identified by a string of three random numbers.

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

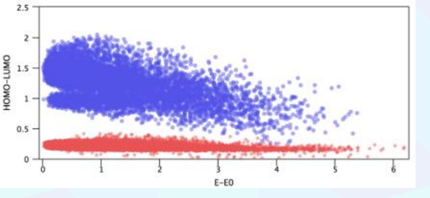
Other Statistical Learning Projects

- Metastabilities of binary compounds, at first considering 5 structures:

Rocksalt CrB CsCl NiAs Zincblende



- Determine the best, correlation-consistent basis functions from a pool of 10,000 Gaussians
- Subgroup discovery algorithms: find structure in big data and analyze what is behind.



Drawing Causal Inference from Big Data (Scientific Insight) -- can we trust a prediction?

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(*)

Judea Pearl

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

1. $d \rightarrow P$: ***P*** "listens" to ***d***
2. $A \rightarrow d$ and $A \rightarrow P$: There is no direct connection between ***d*** and ***P***, but ***d*** and ***P*** both "listen" to a third "actuator"
3. $P \rightarrow d$: ***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) -- can we trust a prediction?

NOVEL MATERIALS DISCOVERY

LASSO has provided us with an equation for the quantitative energy difference:

$$\Delta E = 0.108 \frac{\mathrm{EA(B)} - \mathrm{IP(B)}}{r_p(\mathrm{A})^2} + 1.790 \frac{|r_s(\mathrm{A}) - r_p(\mathrm{B})|}{\exp(r_s(\mathrm{A}))} + 3.766 \frac{|r_p(\mathrm{B}) - r_s(\mathrm{B})|}{\exp(r_d(\mathrm{A}))} - 0.0267$$

This is an equation, not a scientific law:

Case #2:

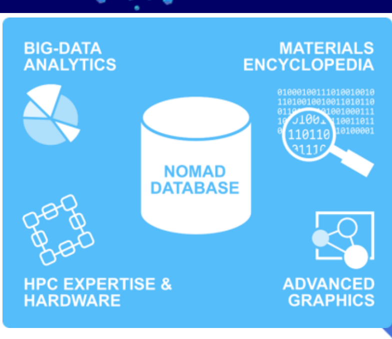
Nuclear numbers $Z_{\mathrm{A}}, Z_{\mathrm{B}}$ ⟷ our descriptor

many-body Hamiltonian ⟶ energy differences

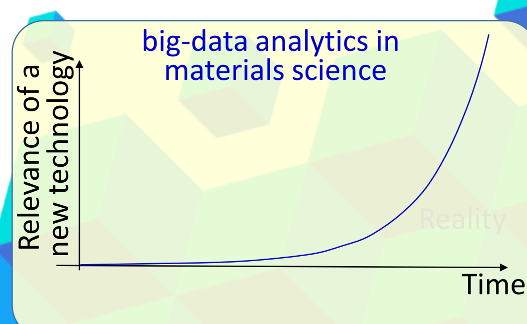
a mapping exists, even a physical intuition exist, but ΔE does not listen directly to the descriptor (intricate causality)

The NOMAD Laboratory

A European Centre of Excellence



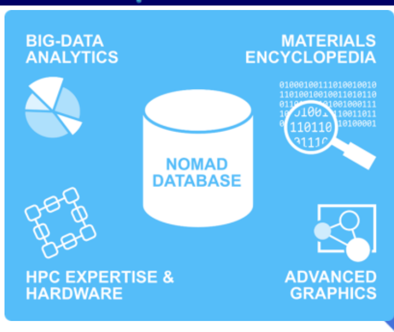
NOMAD develops a *Materials Encyclopedia* and *Big-Data Analytics* and *Advanced Graphics Tools* for materials science and engineering.



The amount of different materials is huge. However, the number of materials that exhibit a certain function, is rather small, i.e. **the space of chemical compounds is sparsely populated**.

We need to develop *domain-specific* compressed-sensing and machine-learning tools.

The NOMAD Laboratory
A European Centre of Excellence



NOMAD develops a *Materials Encyclopedia* and *Big-Data Analytics* and *Advanced Graphics Tools* for materials science and engineering.

The amount of different materials is huge. However, the number of materials that exhibit a certain function, is rather small, i.e. **the space of chemical compounds is sparsely populated**.

We need to develop *domain-specific* compressed-sensing and machine-learning tools.