Making The Data Revolution Happen
Turning Billions of Data from Computational Materials Science into Knowledge by AI

A New Materials-Research Paradigm
History of Paradigms in Materials Science and Engineering

1st PARADIGM
Empirical Science
- Experiments

2nd PARADIGM
Theoretical Science
- Laws of classical mechanics, electrodynamics, etc.

3rd PARADIGM
Computational Science, Simulations
- Density-functional theory and beyond; molecular dynamics

4th PARADIGM
Big-Data-Driven Science
- Detection of patterns and anomalies in Big Data; artificial intelligence; etc.

1600 1950 2010
Jim Gray (Jan. 11, 2007): The 4th Paradigm, Data Intensive Discovery, edited by Hey, Tansley, and Tolle, Microsoft Research

1st PARADIGM
Empirical Science
Experiments
Laws of classical mechanics, electrodynamics, etc.

2nd PARADIGM
Theoretical Science
Density-functional theory and beyond; molecular dynamics
Detection of patterns and anomalies in Big Data; artificial intelligence; etc.

3rd PARADIGM
Computational Science, Simulations

4th PARADIGM
Big-Data-Driven Science

Learning from “Big” Data:
Very Many Methods and Concepts,

Very Interdisciplinary

Big-Data Analytics
Cheminformatics
Intelligent Systems
Robotics

Bioinformatics
Subgroup Discovery

Signal Processing
Data Mining

Computer Science
Clustering

Artificial Intelligence
(learning from experience)

Machine Learning

Data Compression
Neural Networks; Support Vector Machines;
Kernel Ridge Regression; Compressed Sensing; etc.
Big Data of materials has structure and patterns (correlations).

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 data quality).

The NOMAD Laboratory
A European Centre of Excellence
1) Conversion of a Greenhouse Gas into Fuels and ...
2) An Exciton in Lithium Fluoride - Where is the electron?

The NOMAD Repository accepts /requests in- and output files of all important codes. Currently, the NOMAD Repository contains > 50 million total-energy calculations.
The NOMAD Repository and Archive
>50 Mio. Total-Energy Calculations

90% of the VASP files are from
AFLOW OQMD
S. Curtarolo C. Wolverton
Materials Project
G. Ceder K. Persson

The NOMAD Repository accepts requests in- and output files of all important codes. Currently, the NOMAD Repository contains >50 million total-energy calculations.

The NOMAD Laboratory
A European Centre of Excellence

NOMAD Repository
Keeps data for at least 10 years
Open access (if wanted with delay)
Provides DOIs (makes data citable)

https://repository.nomad-coe.eu
also described at youtube.com

40 important codes are used in computational materials science. The heterogeneity could hardly be worse.
The Big-Data Challenge

Big Data of materials has structure and patterns (correlations).

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 data quality).

Max Planck Society

Fig. 1. Historical evolution of the predicted equilibrium lattice parameter for silicon.

All calculations within framework. Values from 15, 16 are compared with (i) predictions from the different codes used in this study (2016 data points, magnified in or calculations with lower numerical settings) and (ii) the 0 K and point effects (red line).

Veracity Has Two Components: Precision and Accuracy

-- Si Lattice Parameter Calculated with DFT-PBE --

Historical evolution of the predicted equilibrium lattice parameter for silicon

Science 351, aad3000 (2016)
Veracity Has Two Components: Precission and Accuracy
-- Si Lattice Parameter Calculated with DFT-PBE --

Historical evolution of the predicted equilibrium lattice parameter for silicon

Influence of pseudo-potential, basis set, k-mesh and other numerical approximations

Veracity: Precision

Science 351, aad3000 (2016)

Reproducibility in density functional theory calculations of solids

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

1 University of Antwerp, 2 University of Antwerp, 3 University of Nijmegen, 4 University of Antwerp, 5 wien2k team, Austria
This study was already discussed by Volker Blum. It identified problems with certain pseudopotentials, e.g. O. This is “only” a test for 71 elemental solids. We need analogous work for defects and surfaces.


Veracity Has Two Components: Precision and Accuracy -- Si Lattice Parameter Calculated with DFT-PBE --

Historical evolution of the predicted equilibrium lattice parameter for silicon

Influence of the DFA, relativistic effects, Born-Oppenheimer A.

Science 351, aad3000 (2016)
Perdew’s Dream: Jacob’s Ladder in Density-Functional Theory

The exchange-correlation functional

our favorite
unoccupied $\psi_i(r)$, EX + cRPA, as given by ACFD
occupied $\psi_i(r)$, hybrids (B3LYP, PBE0, HSE, ...)
$\tau(r)$, meta-GGA (e.g., SCAN)
$\nabla n(r)$, Generalized Gradient Approximation
$n(r)$, Local-Density Approximation

$\tau(r)$: Kohn-Sham kinetic-energy density
EX: exact exchange:
cRPA: random-phase approximation for correlation
ACFD: adiabatic connection fluctuation dissipation theorem

$E_x = -\frac{1}{2} \sum_{\text{occ}} \iint dr \, dr' \frac{\psi_n^*(r) \psi_m(r)}{|r-r'|} \psi_m^*(r') \psi_n(r')$

Bohm, Pines (1953); Gell-Mann, Brueckner (1957); Gunnarsson, Lundqvist (1975, 1976); Langreth, Perdew (1977);

The Big-Data Challenge

Big Data of materials has structure and patterns (correlations).

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 data quality).

Max Planck Society

Big-Data Analytics
The Big-Data Challenge

Big Data of materials has structure and patterns (correlations).

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 data quality).

Statistical Learning by Kernel Ridge Regression

We have data \( \{P_i\} \) at “coordinates” \( \{d_i\} \), \( d_i = \) set of descriptive parameters (descriptor)

\[
P_i \overset{\dagger}{=} P(d_i) = \sum_{k=1}^{N} c_k K(d_i, d_k)
\]

Linear regression: \( K(d_i, d_k) = d_i \cdot d_k \) \hspace{1cm} \( P(d_i) = d_i \cdot c^* \)

Polynomial kernel \( K(d_i, d_k) = (d_i \cdot d_k + c)^d \)

Gaussian kernel \( K(d_i, d_k) = \exp\left(-\sum_j (d_i - d_k)^2 / 2\sigma_j^2\right) \)

More data means a better representation. – Do we learn anything?

Will we ever have enough data?
Statistical Learning by Kernel Ridge Regression

We have data \( \{P_i\} \) at “coordinates” \( \{d_i\} \) \( d_i \) = set of descriptive parameters (descriptor)

\[
P_i \overset{!}{=} P(d_i) = \sum_{k=1}^{N} c_k K(d_i, d_k)
\]

Linear regression: \( K(d_i, d_k) = d_i \cdot d_k \quad P(d_i) = d_i \cdot c^* \)

Polynomial kernel \( K(d_i, d_k) = (d_i \cdot d_k + c)^d \)

Gaussian kernel \( K(d_i, d_k) = \exp\left(-\sum_j (d_i - d_k)^2 / 2\sigma_j^2\right) \)

? More data means a better representation? – Do we learn anything? \( \textbf{Will we ever have enough data?} \)

We have many subgroups in materials science. In these subgroups we typically don’t have big data. We are looking for needles in a haystack.

Create Big Data of Materials -- the SISSO approach --

SISSO (sure independence screening and selection operator) creates \textit{big data of materials}: billions of data that characterize \( N \) different materials.

Then it identifies structure in these big data by analyzing which of this data is relevant for the property of interest
Building Maps of Materials
(Role Models: Periodic Table, Ashby Plots)

Crystal-structure prediction
- Octet binarics (ZB vs. RS)
- $\text{In}_x\text{Al}_y\text{Ga}_z\text{O}_3$ ($x+y+z=2$)
- Perovskites (Goldschmidt tolerance factor)

Property classification:
- Topological insulators

Max Planck Society

Activation of $\text{CO}_2$ at transition-metal oxides and carbides

Property classification:
- Metal vs. insulator

Topological Insulators
(Quantum Spin-Hall Insulators)

conduction band

Trivial insulator

valence band
Topological Insulators
(Quantum Spin-Hall Insulators)


Topological transition

\( \xi_0 \) \( \rightarrow \) \( \xi \)

Spin-orbit coupling, alloying, strain, electric field

2D topological insulators: No backscattering in edge states.
Promising materials in spintronics applications.
Compounds functionalized with F, Cl, Br, I are represented by diamonds, squares, circles and triangles, respectively.

Metals: green,

Functionalization dependent-TIs: red,

Functionalization independent-TIs: blue,

Trivial insulators: white/grey.
Compounds functionalized with F, Cl, Br, I are represented by diamonds, squares, circles and triangles, respectively.

- Metals: green,
- Functionalization dependent-TIs: red,
- Functionalization independent-TIs: blue,
- Trivial insulators: white/grey.

The identified descriptors only depend on properties of the free atoms!

The descriptor identifies materials that do not follow commonly believed rules of thumb.
Turning Greenhouse Gases into Useful Chemicals and Fuels (work in progress)

Formic acid
Formaldehyde
Methanol
Methane

We need an efficient catalyst!

Max Planck Society

From the talk by Weixue Li

Catalysis, Surface Science, Theory
Subgroup discovery reveals two classes of surface terminations (ionic vs covalent)

SISSO provides different descriptors for the two subgroups of CO$_2$ adsorption energy, O-C-O angle, and C-O bond length.

From the talk by Weixue Li
**(Very) Basic Understanding of Heterogenous Catalysis**

**Sabatier:** The interaction between adsorbed reactants and the substrate should be just right: Neither too strong nor too weak.

**Brønsted–Evans–Polanyi:** The reaction barrier scales linearly with the adsorption energy.

A catalysis is “a living” system it changes all the time.

The catalytically active system is (typically) only established after the catalyst material is exposed to reactive conditions (induction period).

---

**Modelling Adsorption at Surfaces and Catalysis -- the position of the $d$-band --**

The position of the $d$ band determines the strength of the chemical bond

*Grimley, Newns, Anderson, Hammer, Norskov*

Modelling Adsorption at Surfaces and Catalysis

-- scaling relations --

Adsorption energies of molecules are proportional to the adsorption energy of atoms. (J. Norskov et al.)

For example, the adsorption energy of CH is proportional to that of C, and the adsorption energy of CO is also proportional to the adsorption energy of C, etc.

Presently the preferred approach. It works better than the $d$-band model.
Still, not good enough.

Prediction Errors: The Materials-Science Challenge Is Different to That of Standard ML

A “75% chance of being right” in the description of the hay is not good enough. RMSE is not a good measure; it mainly refers to the hay but not to the needles.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

$\hat{y}_i$ = predicted value
$y_i$ = true value
A “75% chance of being right” in the description of the hay is not good enough. RMSE is not a good measure; it mainly refers to the hay but not to the needles.

Adsorption energies at “single atom” surface dopants and at AB alloys

RMSE
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 and structural compounds is sparsely populated.
The Novel Materials Discovery (NOMAD) Laboratory maintains the largest Repository for input and output files of all important computational materials science codes.

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

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

NOMAD Scope and Overview
Data is a crucial raw material of the 21st century.
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 and structural compounds is sparsely populated.