Where is condensed-matter theory going?

1st PARADIGM
Empirical Science

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 patterns and anomalies in Big Data; artificial intelligence; etc.

NOMAD

1600 1950 2010
Not-so-open data

Too many funding agencies resist supporting the sharing of data and too many research communities struggle with its practicalities. The result is empty rhetoric and slow science.

Everyone agrees that there are good reasons for having open data. It speeds research, allowing others to build promptly on results. It improves replicability. It enables scientists to test whether claims in a paper truly reflect the whole data set. It helps them to find incorrect data. And it improves the attribution of credit to the data’s originators. But who will pay? And who will host?

Only rarely does a research-funding agency step up to both of these plates. Examples include NASA and the US National Institutes of Health, and the European Bioinformatics Institute. The National Natural Science Foundation of China has ambitions to host the outputs of those that it supports. The European Commission hopes to offer such platforms with its European Open Science Cloud. The UK Data Archive for the social sciences and humanities, and DANS, the Netherlands Institute for Permanent Access to Digital Research Resources, represent other good models of support from governments.

So which fields need to raise their data-access game? Nature suggests that the geodesy and seismology communities should consider revising their current two-year embargo. The microbiome community places great value on open data but, as a relatively young field, is struggling to establish standards.

Thumbs up for two communities that are making progress in this realm. In pathogen genomics, the authors of the Zika virus genome papers we publish in this issue (see pages 401, 406 and 411) made the sequences openly available as soon as they were generated.

Credit should also be given to palaeontologists in their pursuit of an open strategy for 3D data. A recent paper, ‘Open data and digital morphology’ (T. G. Davies et al. Proc. R. Soc. B 284, 20170194; 2017), proposes


Data sharing in materials discovery

Your concerns over impediments to data sharing (Nature 546, 327; 2017) are no longer an issue in computational materials science. This is because the Novel Materials Discovery project (NOMAD; https://nomad-coe.eu) has stimulated a cultural shift in attitudes towards open data as a result of the valuable knowledge that has emerged from data mining since early 2014.

NOMAD provides open access to input and output files from the field’s major data collections (http://aflowlib.org; http://oqmd.org; https://materialproject.org), together with those of individual researchers and groups. It contains more than 40 million total-energy calculations, corresponding to billions of core processing hours by high-performance computers globally.

NOMAD hosts the data for a decade or more, offers digital object identifiers (DOIs) to make data citable, and provides services such as an encyclopedia, big-data analytics tools and advanced graphics.

Hosting the data is expensive, so funding agencies need to step in. However, costs are negligible compared with those of data creation and long-term storage.
The FAIR Concept for Big-Data-Driven Materials Science

C. Draxl and M. Scheffler, MRS Bulletin, Sep. 2018

Findable
Accessible
Interoperable
Re-usable


51.786.061 open-access calculations

https://Repository.NOMAD-CoE.eu
Just four years back ...
# Materials data and their structure

<table>
<thead>
<tr>
<th>Level</th>
<th>Properties</th>
<th>Methods</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Atomic positions and nuclear charges, properties of free atoms, symmetry,</td>
<td>Input: definition of material, etc.</td>
<td>10 kB - 10 MB</td>
</tr>
<tr>
<td></td>
<td>temperature (T), pressure (P)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>Total energy, electron density, potential, wavefunctions, optimized geometry</td>
<td>Density-functional theory (DFT) and</td>
<td>10 MB - 10 TB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ab initio molecular dynamics (MD)</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>Excitation, dielectric, Coulomb, phonon spectra, thermal conductivity, etc.</td>
<td>Many-body perturbation theory (MBPT),</td>
<td>1 GB - 10 TB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DF perturbation theory, ab initio MD</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td><strong>Thermoelectric figure of merit</strong>, turn-over frequency of catalyst, efficiency of solar cell, etc. as a function of T and P</td>
<td>Modeling, output derived from levels I-III phenotype</td>
<td>10 kB - 1 MB</td>
</tr>
</tbody>
</table>

*The amount of materials data produced on workstations, compute clusters, and supercomputers is growing exponentially.*

*Most of it is thrown away...*
## Materials data and their structure

<table>
<thead>
<tr>
<th>Level</th>
<th>Properties</th>
<th>Methods</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Atomic positions and nuclear charges, properties of free atoms, symmetry, temperature (T), pressure (P)</td>
<td>Input: definition of material, gene</td>
<td>10 kB - 10 MB</td>
</tr>
<tr>
<td>II</td>
<td>Total energy, electron density, potential, wavefunctions, atomic forces, optimized geometry, elastic constants, etc.</td>
<td>Density-functional theory (DFT) and ab initio molecular dynamics (MD)</td>
<td>10 MB - 10 TB</td>
</tr>
<tr>
<td>III</td>
<td>Excitation energies, electrical conductivity, dielectric screening, matrix elements of Coulomb interaction, etc. optical spectra, phonon spectra, thermal conductivity, etc.</td>
<td>Many-body perturbation theory (MBPT), DF perturbation theory, ab initio MD</td>
<td>1 GB - 10 TB</td>
</tr>
<tr>
<td>IV</td>
<td>Thermoelectric figure of merit, turn-over frequency of catalyst, efficiency of solar cell, etc. as a function of T and P</td>
<td>Modeling, output derived from levels I-III phenotype</td>
<td>10 kB - 1 MB</td>
</tr>
</tbody>
</table>

---

**Why should we compute the same again?**

**Much of the value of high-throughput calculations is wasted without deeper Big-Data driven analysis of the results.**
Where are we aiming at?
Our vision is to draw maps

What are the actuators behind the trends and patterns that are invisible to the human eye?

- Thermal-barrier coatings
- Super-conductors
- Materials for photovoltaics
- Transparent metals
The NOMAD Archive

More than 51 million calculations coming from ...
Making scientific data accessible
You may add items from any of the three categories to your search.
AgFeO₃ - space group 221

**Structure**
- System type: bulk
- Space group: 221
- Structure type: CaO3Ti (Cubic Perovskite)

**Electronic structure**
- Band structure
- DOS
- From calculation 1648548 (GGA - VASP)
- From calculation 383297 (GGA - VASP)

**Methodology**
- Available calculations
  - Functional: 7 GGA
  - Code: 7 VASP

https://encyclopedia.nomad-coe.eu/
Thermal properties

Ge - space group 227

Calculations

- Germanium
- GGA
- LDA
- PHL-aims (484)
  - 232721
  - 240305
  - 245463
  - 248199
  - 247902
  - 334303

Vibrational and thermal properties

Phonon dispersion

Phonon DOS

Frequency (cm⁻¹)
Can we trust computed data?
Reproducibility in density functional theory calculations of solids

Delta test

Compute E(V) using PBE
Do the same with other code
Fit to Burch-Murnaghan equation of state

Quality factor

$$\Delta = \left( \sqrt{\int \frac{\Delta E^2(V) dV}{\Delta V}} \right)$$

Energy

Volume

code 1

code 2
<table>
<thead>
<tr>
<th>Code</th>
<th>Version</th>
<th>Basis</th>
<th>Electron treatment</th>
<th>Δ-value</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exciting</td>
<td>development LAPW+xlo version</td>
<td>all-electron</td>
<td>0 meV/atom</td>
<td>Exciting [10,16]</td>
<td></td>
</tr>
<tr>
<td>FHI-aims</td>
<td>081213</td>
<td>tier2 numerical all-electron (relativistic atomic_zora scalar)</td>
<td>0.1 meV/atom</td>
<td>ASE [2,16]</td>
<td></td>
</tr>
<tr>
<td>WIEN2k</td>
<td>13.1</td>
<td>LAPW/APW+lo all-electron</td>
<td>0.2 meV/atom</td>
<td>S. Cottenier [16]</td>
<td></td>
</tr>
<tr>
<td>FHI-aims</td>
<td>081213</td>
<td>tier2 numerical all-electron (relativistic zora scalar 1e-12)</td>
<td>0.3 meV/atom</td>
<td>ASE [2]</td>
<td></td>
</tr>
<tr>
<td>Quantum ESPRESSO</td>
<td>5.1</td>
<td>plane waves SSSP Accuracy (mixed NC/US/PAW potential library)</td>
<td>0.3 meV/atom</td>
<td>QuantumESPRESSO [12,16]</td>
<td></td>
</tr>
<tr>
<td>Code</td>
<td>Version</td>
<td>Basis</td>
<td>Electron treatment</td>
<td>Δ-value</td>
<td>Authors</td>
</tr>
<tr>
<td>--------------</td>
<td>---------</td>
<td>------------------------------</td>
<td>-------------------------------------------</td>
<td>---------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td><strong>Exciting</strong></td>
<td>development LAPW+xlo version</td>
<td>all-electron</td>
<td>0 meV/atom</td>
<td></td>
<td>[10,16]</td>
</tr>
<tr>
<td><strong>FHI-aims</strong></td>
<td>081213</td>
<td>tier2 numerical atomic_zora scalar</td>
<td>0.1 meV/atom</td>
<td></td>
<td>[2,16]</td>
</tr>
<tr>
<td><strong>WIEN2k</strong></td>
<td>13.1</td>
<td>LAPW/APW+lo</td>
<td>all-electron</td>
<td>0.2</td>
<td>S. Cottenier [16]</td>
</tr>
<tr>
<td><strong>FHI-aims</strong></td>
<td>081213</td>
<td>tier2 numerical ZORA scalar 1e-12</td>
<td>0.3 meV/atom</td>
<td></td>
<td>[2]</td>
</tr>
<tr>
<td><strong>Quantum ESPRESSO</strong></td>
<td>5.1</td>
<td>plane waves</td>
<td>SSSP Accuracy (mixed NC/US/PAW potential library)</td>
<td>0.3</td>
<td>QuantumESPRESSO [12,16]</td>
</tr>
</tbody>
</table>
Back to our vision ...

What are the actuators behind the trends and patterns that are invisible to the human eye?
Big-Data analytics - an example

Phillips - van Vechten problem

Given atoms A and B
What crystal would they form?

rocksalt  zincblende

Big-Data analytics - an example

Structure map of binary semiconductors, obtained with a compressed-sensing algorithm.

Predictions from free neutral atoms A and B

Results can be reenacted at the NOMAD Analytics Toolkit

https://analytics-toolkit.nomad-coe.eu/

Model building

**Training set**
Calculate property $P$ for many materials

**Descriptors**
Build feature space, $d$

**Cross validation**

**Predictions**
Calculate property for test set / new materials

**Learning**
Find function $P(d)$ for training data

*e.g. LASSO*
# Building descriptors

## Primary features

### Free atoms

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP(A), IP(B)</td>
<td>Ionization potential</td>
</tr>
<tr>
<td>EA(A), EA(B)</td>
<td>Electron affinity</td>
</tr>
<tr>
<td>H(A), H(B)</td>
<td>Highest occupied Kohn-Sham level</td>
</tr>
<tr>
<td>L(A), L(B)</td>
<td>Lowest unoccupied Kohn-Sham level</td>
</tr>
<tr>
<td>r_s(A), r_s(B)</td>
<td>Radius at max. of s-like wavefunction</td>
</tr>
<tr>
<td>r_p(A), r_p(B)</td>
<td>Radius at max. of p-like wavefunction</td>
</tr>
<tr>
<td>r_d(A), r_d(B)</td>
<td>Radius at max. of d-like wavefunction</td>
</tr>
</tbody>
</table>

### Dimers

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>H(L)(AA), H(L)(BB), H(L)(AB)</td>
<td>HOMO-LUMO KS gap</td>
</tr>
<tr>
<td>E_b(AA), E_b(BB), E_b(AB)</td>
<td>Binding energy</td>
</tr>
<tr>
<td>d(AA), d(BB), d(AB)</td>
<td>Equilibrium distance</td>
</tr>
</tbody>
</table>
Building descriptors

Full feature space

10000 nonlinear combinations of primary features

+, -, *, /, ^2, ^3, \sqrt{}, \exp{}, ...

Linear relationship \( P(d) = c_d \)

Let the machine choose most relevant descriptors

\[
\min_{c \in \mathbb{R}^M} \|P - c_d\|^2 + \lambda \|c\|_1
\]
Funding from the European Union’s Horizon 2020 research and innovation programme, grant agreement No 676580.