HPC for Computationally and Data-Intensive Problems

Luca M. Ghiringhelli
Fritz-Haber-Institut der MPG, Berlin

SuperMUC-NG
Next-Gen Science Symposium
Munich, November 22, 2018
Data-driven materials science
The Big Picture

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

• About 240,000 *inorganic* materials are known to exist (Springer Materials)
Design of new materials: preparation, synthesis, and characterization is complex and costly.

About 240,000 *inorganic* materials are known to exist (Springer Materials).

Basic properties determined for very few of them.
Design of new materials: preparation, synthesis, and characterization is complex and costly.

About 240,000 inorganic materials are known to exist (Springer Materials).

Basic properties determined for very few of them.

Number of possible materials: practically infinite.
Design of new materials: preparation, synthesis, and characterization is complex and costly.

About 240,000 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.
Design of new materials: preparation, synthesis, and characterization is complex and costly.

About 240,000 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.
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
The NOMAD Laboratory
https://nomad-coe.eu

Archive (code independent)
52M total-energy calculations
[90% coming from
AFLOW (Curtarolo)
OQMD (Wolverton)
Materials Project (Ceder)]
Data analytics: an ideal flow chart

**Training set**
Calculate properties and functions $P_i$, for many materials, $i$
E.g., Density-Functional Theory
Data analytics: an ideal flow chart

**Descriptor**
Find the appropriate descriptor $d_i$, build a table:

| $i$ | $d_i$ | $P_i$ |

**Training set**
Calculate properties and functions $P_i$, for many materials, $i$
E.g., Density-Functional Theory
Find the appropriate descriptor $d_i$, build a table:

| $i$ | $d_i$ | $P_i$ |

Learning

Find the function $P(d)$ for the table. Build a chart for the property

Statistical learning
Data analytics: an ideal flow chart

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

**Fast Prediction**
Calculate properties and functions for new values of $d$
(new materials)

**Descriptor**
Find the appropriate descriptor $d_i$, build a table:

| $i$ | $d_i$ | $P_i$ |

**Learning**
Find the function $P(d)$ for the table.
Build a chart for the property
Statistical learning
Data analytics: an ideal flow chart

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

**Descriptor**
Find the appropriate descriptor $d_i$, build a table:

| $i$ | $d_i$ | $P_i$ |

**Fast Prediction**
Calculate properties and functions for new values of $d$ (new materials)

**Learning**
Find the function $P(d)$ for the table.
Build a chart for the property

Statistical learning
Learning/discovering maps of materials properties. A quantum many-body problem

Octet binaries

$E_{\text{ads}}(\text{CO}_2)$ on oxides

$A_xB_y$ binaries

2D honeycomb materials

Trivial insulator

Topological insulator

Metal

Insulator

Metal
Compressed sensing: the quest for descriptors and predictive models

Structure map with compressed-sensing algorithm, starting from 7 atomic features
Compressed sensing
Aim: finding descriptors and learning predictive models

Ansatz:
\[ P = c_1 d_1 + c_2 d_2 + \ldots + c_n d_n \]

\( P \): property of interest
\( d_1, \ldots d_n \): features, i.e., (nonlinear) functions of primary features (EA, IP, ...)
\( c_1, \ldots c_n \): unknown coefficients \(\Rightarrow\) as few as possible are nonzero
Compressed sensing
Aim: finding descriptors and learning predictive models

Ansatz:
\[ P = c_1 d_1 + c_2 d_2 + \ldots + c_n d_n \]

\( P \): property of interest
\( d_1, \ldots, d_n \): features, i.e., (nonlinear) functions of primary features (EA, IP, ...)
\( c_1, \ldots, c_n \): unknown coefficients => as few as possible are nonzero

d_i: iterative construction with features and operators (+, \times, /, ^2, ...)
Compressed sensing
Aim: finding descriptors and learning predictive models

Ansatz:
\[ P = c_1 d_1 + c_2 d_2 + \ldots + c_n d_n \]

\( P \): property of interest
\( d_1, \ldots d_n \): features, i.e., (nonlinear) functions of primary features (EA, IP, ...)
\( c_1, \ldots c_n \): unknown coefficients => as few as possible are nonzero

\[ \text{argmin}_{c \in \mathbb{R}^M} \| P - Dc \|_2^2 + \lambda \| c \|_0 \]

\[ \frac{\text{IP}(B) - \text{EA}(B)}{r_p(A)^2}, \quad \frac{|r_s(A) - r_p(B)|}{\exp(r_s(A))} \]
SISSO: sure independence screening plus sparsifying operator

$$\arg\min_{c \in \mathbb{R}^M} \|P - Dc\|_2^2 + \lambda \|c\|_0$$

Huge “feature space” (e.g., 100 billion elements) constructed from primary features (e.g., 15 - 25 elements)
SISSO: sure independence screening plus sparsifying operator

$$\arg\min_{c \in \mathbb{R}^M} \| P - Dc \|_2^2 + \lambda \| c \|_0$$

Huge “feature space” (e.g., 100 billion elements) constructed from primary features (e.g., 15 - 25 elements)
SISSO: sure independence screening plus sparsifying operator

\[
\arg\min_{c \in \mathbb{R}^M} \| P - Dc \|_2^2 + \lambda \| c \|_0
\]

Huge “feature space” (e.g., 100 billion elements) constructed from primary features (e.g., 15 - 25 elements)
SISSO: sure independence screening plus sparsifying operator

\[
\underset{c \in \mathbb{R}^M}{\text{argmin}} \| P - Dc \|^2_2 + \lambda \| c \|_0
\]

Huge “feature space” (e.g., 100 billion elements) constructed from primary features (e.g., 15 - 25 elements)

SISSO: sure independence screening plus sparsifying operator

Embarrassingly parallel
+ SIS: independent scalar products of features on property or residual
- partial ranking
+ SO: independent least square regression
- partial ranking
+ outer parallelization for cross validation
- smart strategies needed for matrix storage
SISSO: sure independence screening plus sparsifying operator

Compressed sensing: the quest for descriptors and predictive models

Structure map with compressed-sensing algorithm, starting from 7 atomic features
Charts/maps of materials

argmin(\(\|P - Dc\|^2_2 + \lambda \|c\|_0\))
\(c \in \mathbb{R}^M\)

New cost function to be minimized: overlap of convex domains

\[\hat{c} \equiv \arg \min_c \left( \sum_{i=1}^{M-1} \sum_{j=i+1}^M O_{ij} + \lambda \|c\|_0 \right)\]

Number of data points in the overlap region, as function of selected \(d\)

Iterative generation of feature subspaces
SISSO: metal/nonmetal classification of binary materials

\[ \chi_A^2 \left| 1 - 2x_A \right| - x_A^2 \frac{\chi_B}{\chi_A} \]

- 'x': Atomic fraction
- 'IE': Ionization energy
- '\chi': Electronegativity

SISSO: predicting novel honeycomb (~2D) topological insulators

Data source: high throughput DFT (FHI-aims, Carlos Mera Acosta)
SISSO: predicting novel honeycomb (~2D) *topological insulators*

![Graph showing the classification of materials into trivial insulators, functionalization-dependent topological insulators, functionalization-independent topological insulators, metals, and functionalization-dependent topological insulators. The graph plots different elements (F, Cl, Br, I) against a parameter involving electron affinity, ionization potential, and other properties.](image-url)
Perovskites’ stability: Improving on Goldschmidt Tolerance Factor

$ABX_3$

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

Ionic radius

Goldschmidt* stable perovskites:

$t = 0.825 \quad t = 1.059$

accuracy 74%

Perovskites’ stability: Improving on Goldschmidt Tolerance Factor

\[ t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)} \]  
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 / \( \mu \) = Octahedral factor

Goldschmidt* stable perovskites: accuracy 74%

\[ t = 0.825 \quad t = 1.059 \]

Our stable perovskites: accuracy 92%

\[ \tau = 4.18 \]

Perovskites’ stability: Improving on Goldschmidt Tolerance Factor

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

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

1 / \mu = Octahedral factor

Goldschmidt* stable perovskites: accuracy 74%

\[ t = 0.825 \quad t = 1.059 \]

Our stable perovskites: accuracy 99%

\[ \tau = 3.31 \quad \tau = 4.18 \quad \tau = 5.92 \]

Perovskites’ stability: Improving on Goldschmidt Tolerance Factor

\[ t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)} \] 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 / \mu = Octahedral factor

Goldschmidt* stable perovskites: accuracy 74%

Our stable perovskites: accuracy 100%

Continuous property

- Adsorption energy of O on metal-oxide surfaces
- Adsorption energy and OCO angle of adsorbed CO$_2$ on metal-oxide surfaces
- Adsorption energy of metal atoms on metal-alloys surfaces

Features: atoms (of the surface) and pristine surface

Classification

- Tetradyomite 5-component 3d topological insulators (vs trivial insulators) arXiv:1808.04733

Features: atoms
Convolutional neural networks for (local) crystal-structure recognition
Acknowledgements

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

SISSO and metal/insulator proof of concepts
Runhai Ouyang, Emre Ahmetcik, Stefano Curtarolo

Topological Insulators
Carlos Mera Acosta, Adalberto Fazzio, Runhai Ouyang, Christian Carbogno

Perovskites
Christopher J. Bartel, Christopher Sutton, Bryan R. Goldsmith, Runhai Ouyang, Charles B. Musgrave

Convolutional neural networks for structure recognition
Angelo Ziletti, Andreas Leitherer, Devinder Kumar

NOMAD PIs
Angel Rubio (MPSD Hamburg), Risto Nieminen (Aalto Univ. Helsinki), Francesc Illas (Univ. Barcelona), Daan Frenkel (Univ. Cambridge), Claudia Draxl (HU Berlin), Alessandro De Vita (Kings College London), Kristian Thygesen (DTU Lyngby); Kimmo Koski (CSC Helsinki), Stefan Heinzel (MPSCD Garching), Jose Maria Cela (BSC Barcelona), Dieter Kranzlmüller (LRZ Munich); Ciaran Clissman (Pintail Dublin)

All the above
Matthias Scheffler