(Big-)data analytics

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

Hands-On DFT and Beyond: Frontiers of Advanced Electronic Structure and Molecular Dynamics Methods
Peking University, Beijing, China, July 30th to August 10th, 2018
Charts/maps of materials
Charts/maps of materials

Octet binaries

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

$|E_{\text{ads}}|$
Building maps of materials properties
A quantum many-body problem

From the Hamiltonian of a physical system, in principle we can derive all properties (observables).
Building maps of materials properties
A quantum many-body problem

From the Hamiltonian of a physical system, in principle we can derive all properties (observables). But in practice, the Hamiltonian is often not the starting point.
Building maps of materials properties
A quantum many-body problem

From the Hamiltonian of a physical system, in principle we can derive all properties (observables). But in practice, the Hamiltonian is often not the starting point.

For instance, given a class of chemical compositions (e.g., via prototype formula, such as $ABX_3$):
From the Hamiltonian of a physical system, in principle we can derive all properties (observables). But in practice, the Hamiltonian is often not the starting point.

For instance, given a class of chemical compositions (e.g., via prototype formula, such as ABX$_3$):

- what is the most stable crystal structure of each material in the class?
From the Hamiltonian of a physical system, in principle we can derive all properties (observables). But in practice, the Hamiltonian is often not the starting point.

For instance, given a class of chemical compositions (e.g., via prototype formula, such as ABX$_3$):
- what is the most stable crystal structure of each material in the class?
- which materials are metals / topological insulators / superconductors?
- which material has the highest melting point?
- which materials has a surface optimal for catalysing some chemical reaction?
Design of new materials: preparation, synthesis, and characterization is complex and costly.
The Big Picture

- 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
The Big Picture

- 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
The Big Picture

- 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

  \( \Rightarrow \) 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
Applications:
- prediction of ground-state (and metastable) crystal structure(s) from the composition
- predicting whether a (binary) material is a metal or insulator
- predicting novel perovskite materials
- predicting properties of transparent conducting oxides
- classifying pristine and defected atomic structures
Methods featured in this lecture:

**Compressed sensing** for building low-dimensional maps

*Intermezzo: crow-sourcing data-driven materials science*

*The NOMAD-kaggle competition*

**Subgroup discovery** for finding outstanding subsets of data

**Convolutional neural networks** for structure (image) recognition

Applications:
- prediction of ground-state (and metastable) crystal structure(s) from the composition
- predicting whether a (binary) material is a metal or insulator
- predicting novel perovskite materials
- predicting properties of transparent conducting oxides
- classifying pristine and defected atomic structures
Methods featured in this lecture:

**Compressed sensing** for building low-dimensional maps

*Intermezzo: crow-sourcing data-driven materials science*

*The NOMAD-kaggle competition*

**Subgroup discovery** for finding outstanding subsets of data

**Convolutional neural networks** for structure (image) recognition

**Outline**

Applications:

- prediction of ground-state (and metastable) crystal structure(s) from the composition
- predicting whether a (binary) material is a metal or insulator
- predicting novel perovskite materials
- predicting properties of transparent conducting oxides
- classifying pristine and defected atomic structures
(Big-)data analysis: a flow chart

- **Training set**
  - Calculate properties and functions $P$, for many materials, $i$
  - **Density-Functional Theory**
(Big-)data analysis: a flow chart

**Training set**
Calculate properties and functions $P_i$, for many materials, $i$

**Density-Functional Theory**

**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, $i$
Density-Functional Theory

Descriptor
Find the appropriate descriptor $d_i$,
build a table:
| $i$ | $d_i$ | $P_i$ |

Learning
Find the function $P^{SL}(d)$ for the table;
do cross validation.
Statistical learning
(Big-)data analysis: a flow chart

**Training set**
Calculate properties and functions $P$, for many materials, $i$

**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 $P^{SL}(d)$ for the table; do cross validation.

**Statistical learning**
Training set
Calculate properties and functions $P$, for many materials, $i$

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 $P_{SL}(d)$ for the table; do cross validation.

Statistical learning

(Big-)data analysis: a flow chart
(Big-)data analysis: a flow chart

**Training set**
Calculate properties and functions $P$, for many materials, $i$

**Density-Functional Theory**

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

**What’s “big”, then?**

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

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

**Learning**
Find the function $P_{SL}(d)$ for the table; do cross validation.

**Statistical learning**
Training set
Calculate properties and functions $P$, for many materials, $i$
Density-Functional Theory

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

What’s “big”, then?
- Volume
- Velocity
- Variety
- Veracity issue

Descriptor
Find the appropriate descriptor $d_i$, build a table:
| $i$ | $d_i$ | $P_i$ |

Learning
Find the function $P^{SL}(d)$ for the table; do cross validation.
Statistical learning

(Big-)data analysis: a flow chart
82 octet AB binary compounds

We have a dream

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

An example: predicting crystal structures from the composition
An example: predicting crystal structures from the composition

82 octet AB binary compounds

- Rock salt
- Rock salt/Zinc blende
- Zinc blende
An example: predicting crystal structures from the composition

82 octet AB binary compounds

An example: predicting crystal structures from the composition

82 octet AB binary compounds

The descriptor proposed by Phillips and van Vechten in 1969-70 depends on:
- lattice parameter
- electrical conductivity

An example: predicting crystal structures from the composition

82 octet AB binary compounds

Ansatz: atomic features
- HOMO
- LUMO
- Ionization Potential
- Electron Affinity
- Radius of valence s orbital
- Radius of valence p orbital
- Radius of valence d orbital
- ... ?


We have a dream

Proof of Concept: Descriptor for the Classification

"Zincblende/Wurtzite or Rocksalt?"

An example: predicting crystal structures from the composition
Primary (atomic) features

example: Sn (Tin)
Primary (atomic) features

example: Sn (Tin)

KS levels [eV]

LUMO
Valence p (HOMO)
Valence s

KS levels [eV]
Primary (atomic) features

example: Sn (Tin)

Radial probability densities

Valence s
Valence p

Average radius
Radius @ max
Turning point

Valence s
Valence p
LUMO
Valence p (HOMO)

KS levels [eV]
K S l e v e l s [Å]

KS levels [eV]

Valence s
Valence p
LUMO
Valence p (HOMO)

KS levels [eV]

Valence s
Valence p
LUMO
Valence p (HOMO)

KS levels [eV]

Valence s
Valence p
LUMO
Valence p (HOMO)

KS levels [eV]
An example: predicting crystal structures from the composition

82 octet AB binary compounds

Ansatz: atomic features

- HOMO
- LUMO
- Ionization Potential
- Electron Affinity
- Radius of valence $s$ orbital
- Radius of valence $p$ orbital
- Radius of valence $d$ orbital
- ... ?

$E(\text{Rock salt}) - E(\text{Zinc blende})$

Aim: finding descriptors and learning predictive models

Ansatz:

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

Where

- \( P \) is the property of interest
- \( d_1, \ldots, d_n \) are candidate features, i.e., nonlinear functions of primary features (EA, IP, ...)
- \( c_1, \ldots, c_n \) are unknown coefficients, with the extra constraint that these (nonzero) coefficients should be as few as possible.
Aim: finding descriptors and learning predictive models
Ansatz:
\[ P = c_1 d_1 + c_2 d_2 + \ldots + c_n d_n \]
Where
- \( P \) is the property of interest
- \( d_1, \ldots, d_n \) are candidate features, i.e., nonlinear functions of primary features (EA, IP, ...)
- \( c_1, \ldots, c_n \) are unknown coefficients, with the extra constraint that these (nonzero) coefficients should be as few as possible.

With a foreword on dimensionality reduction
Pearson, K. "On Lines and Planes of Closest Fit to Systems of Points in Space". Philosophical Magazine 2, 559 (1901)
Orthonormal transformation of coordinates, converting a set of (possibly) linearly correlated coordinates into a new set of linearly uncorrelated (called principal or normal) components, such that the first component has the largest variance and each subsequent has the largest variance constrained to being orthogonal to all the preceding components.
Principal component analysis

Pearson, K. "On Lines and Planes of Closest Fit to Systems of Points in Space". Philosophical Magazine 2, 559 (1901)

Orthonormal transformation of coordinates, converting a set of (possibly) linearly correlated coordinates into a new set of linearly uncorrelated (called principal or normal) components, such that the first component has the largest variance and each subsequent has the largest variance constrained to being orthogonal to all the preceding components.

Ansatz: atomic features

- Valence number  
- Energy of valence s orbital  
- Energy of valence p orbital  
- Radius of valence s orbital  
- Radius of valence p orbital 

\[ r_s, r_p, E_s/\sqrt{Z_v}, E_p/\sqrt{Z_v}, \]

for A and B atoms

Saad, …, Chelikowsky, and Andreoni, PRB 85, 104104 (2012)
Linear dimensionality reduction: Principal components

Ansatz: atomic features

- Valence number \( Z_v \)
- Energy of valence s orbital \( E_s \)
- Energy of valence p orbital \( E_p \)
- Radius of valence s orbital \( r_s \)
- Radius of valence p orbital \( r_p \)

\[ r_s, r_p, E_s/\sqrt{Z_v}, E_p/\sqrt{Z_v}, \]
for A and B atoms

linearly uncorrelated (called principal or normal) components, such that the first component has the largest variance and each subsequent has the largest variance constrained to being orthogonal to all the preceding components

Saad, …, Chelikowsky, and Andreoni, PRB 85, 104104 (2012)
Linear dimensionality reduction: Principal components

Ansatz: atomic features

- Valence number $Z_v$
- Energy of valence $s$ orbital $E_s$
- Energy of valence $p$ orbital $E_p$
- Radius of valence $s$ orbital $r_s$
- Radius of valence $p$ orbital $r_p$

$r_s, r_p, E_s / \sqrt{Z_v}, E_p / \sqrt{Z_v}$, for A and B atoms

What's on the axes?

Linear combination of (possibly all) the initial dimensions

Saad, …, Chelikowsky, and Andreoni, PRB 85, 104104 (2012)
Compressed sensing: the quest for descriptors and predictive models

82 octet AB binary compounds

Ansatz: atomic features

- HOMO
- LUMO
- Ionization Potential
- Electron Affinity
- Radius of valence $s$ orbital
- Radius of valence $p$ orbital
- Radius of valence $d$ orbital
- Thousands to billions of non-linear functions of the above

$P = c_1 d_1 + c_2 d_2 + \ldots + c_n d_n$

$E(\text{Rock salt}) - E(\text{Zinc blende})$

- Rock salt
- RS / ZB
- Zinc blende

We have a dream

B) Proof of Concept: Descriptor for the Classification "Zincblende/Wurtzite or Rocksalt?"
Compressed sensing: the quest for descriptors and predictive models

82 octet AB binary compounds

Ansatz: atomic features
- HOMO
- LUMO
- Ionization Potential
- Electron Affinity
- Radius of valence s orbital
- Radius of valence p orbital
- Radius of valence d orbital
- Thousands to billions of non-linear functions of the above

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

Symbolic Regression

\[ E(\text{Rock salt}) - E(\text{Zinc blende}) \]
Systematic construction of the feature space

+ Radius 1 Radius 2

KS level 1 | x - y |
KS level 2

(Linear) dimensionality reduction: principal components
Systematic construction of the feature space
Systematic construction of the feature space

$$\exp(x) \quad (x)^n$$

(Linear) dimensionality reduction: principal components
Systematic construction of the feature space: EUREQA


T. Müller et al. PRB 89 115202 (2014): Data: ~1000 amorphous structures of 216 Si atoms (saturated)

Property: hole trap depth

\[
\frac{\min(1.66355, a) \max(5.37551, c) - f - bd}{g - h \max(3.42929, e)},
\]

Descriptor (candidates: 242)

a The largest distance between a H atom and its nearest Si neighbor
b The shortest distance between a Si atom and its sixth-nearest Si neighbor
c The maximum bond valence sum on a Si atom
d The smallest value for the fifth-smallest relative bond length around a Si atom
e The fourth-shortest distance between a Si atom and its eighth-nearest neighbor
f The second-shortest distance between a Si atom and its fifth-nearest neighbor
g The third-shortest distance between a Si atom and its sixth-nearest neighbor
h The H-Si nearest-neighbor distance for the hydrogen atom with the fourth-smallest difference between the distances to the two Si atoms nearest to a H atom
Compressed sensing: the quest for descriptors and predictive models

82 octet AB binary compounds

Ansatz: atomic features

- HOMO
- LUMO
- Ionization Potential
- Electron Affinity
- Radius of valence $s$ orbital
- Radius of valence $p$ orbital
- Radius of valence $d$ orbital
- Thousands of non-linear functions of the above

$E(\text{Rock salt}) - E(\text{Zinc blende})$

$P = c_1 d_1 + c_2 d_2 + ... + c_n d_n$

$\text{argmin}_{c \in \mathbb{R}^M} \| P - Dc \|^2_2 + \lambda \| c \|_0$
Compressed sensing: the quest for descriptors and predictive models

Ideal method: regression with $\ell_0$ regularization

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

Optimal solution
Non-polynomial complexity
Small # columns in $D$
Compressed sensing: the quest for descriptors and predictive models

Ideal method: regression with $\ell_0$ regularization

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

Optimal solution
Non-polynomial complexity
Small # columns in $D$

$\| c \|_0$ # of nonzero elements of $c$
$\| c \|_2$ Euclidean. Square root of sum of squares of the elements of $c$
Compressed sensing: the quest for descriptors and predictive models

Ideal method: regression with $\ell_0$ regularization

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

Optimal solution
Non-polynomial complexity
Small # columns in $D$

For matrices $D$ with uncorrelated columns: LASSO

$$\arg\min_{c \in \mathbb{R}^M} \| P - Dc \|_2^2 + \lambda \| c \|_1$$

(Possibly) optimal solution
Convex optimization
Moderate # columns in $D$

$\| c \|_0$ # of nonzero elements of $c$
$\| c \|_2$ Euclidean. Square root of sum of squares of the elements of $c$
$\| c \|_1$ “Manhattan”. Sum of absolute values of the elements of $c$
Lattice Anharmonicity and Thermal Conductivity from Compressive Sensing of First-Principles Calculations

Fei Zhou (周非)

Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California 94550, USA

Weston Nielson, Yi Xia, and Vidvuds Ozoliņš

Department of Materials Science and Engineering, University of California, Los Angeles, California 90095-1595, USA

(Received 22 April 2014; published 27 October 2014)

Compressed modes for variational problems in mathematics and physics

Vidvuds Ozoliņša,b, Rongjie Laib,1, Russel Caflischc,1, and Stanley Osher1,2

Departments of aMaterials Science and Engineering, and bMathematics, University of California, Los Angeles, CA 90095-1555; and cDepartment of Mathematics, University of California, Irvine, CA 92697-3875

Contributed by Stanley Osher, October 8, 2013 (sent for review September 3, 2013)

PHYSICAL REVIEW B 87, 035125 (2013)

Compressive sensing as a paradigm for building physics models

Lance J. Nelson and Gus L. W. Hart

Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602, USA

Fei Zhou (周非) and Vidvuds Ozoliņš

Department of Materials Science and Engineering, University of California, Los Angeles, California 90095, USA

(Received 26 June 2012; revised manuscript received 26 September 2012; published 18 January 2013)
Compressed sensing: the quest for descriptors and predictive models

\[ \text{argmin}_{c \in \mathbb{R}^M} \| P - Dc \|_2^2 + \lambda \| c \|_1 \]  

(Possibly) optimal solution
Convex optimization
Moderate # columns in \( D \)

\[ \| c \|_1 \]  

“Manhattan”. Sum of absolute values of the elements of \( c \)
Compressed sensing: the quest for descriptors and predictive models

When there are high correlations, LASSO+$\ell_0$ (LMG et al. PRL 2015):
- use LASSO with lambda in order to “switch on” few tens features (say 30-50)
- perform $\ell_0$ regularization, i.e., for 1,2,3D solution, enumerate all 1- 2- 3-tuples and find the best fitting tuple.

\[
\text{argmin}_{\mathbf{c} \in \mathbb{R}^M} \left\| \mathbf{P} - \mathbf{Dc} \right\|_2^2 + \lambda \left\| \mathbf{c} \right\|_1
\]

(Possibly) optimal solution
Convex optimization
Moderate # columns in $\mathbf{D}$

$\left\| \mathbf{c} \right\|_1$ “Manhattan”. Sum of absolute values of the elements of $\mathbf{c}$
Compressed sensing: the quest for descriptors and predictive models

82 octet AB binary compounds

Ansatz: atomic features

- HOMO
- LUMO
- Ionization Potential
- Electron Affinity
- Radius of valence s orbital
- Radius of valence p orbital
- Radius of valence d orbital
- Billions of non-linear functions of the above

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

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

\( E(\text{Rock salt}) - E(\text{Zinc blende}) \)
Compressed sensing: the quest for descriptors and predictive models

From orthogonal matching pursuit ....

Local-optimum solution
Huge # columns in D

feature
Residual
Property
feature
Compressed sensing: the quest for descriptors and predictive models

From orthogonal matching pursuit ....

... to Sure Independence Screening + Sparsifying Operator (SISSO)

Local-optimum solution
Huge # columns in D

Proxy of
global-optimum solution
Huge # columns in D

R. Ouyang et al. PRM 2, 083802 (2018), published 7 August 2018

Zincblende/Wurtzite or Rocksalt?

Compressed sensing: the quest for descriptors and predictive models
We have a dream

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

Compressed-sensing-based model identification:
Shares concepts with
- Dimensionality reduction (but supervised, i.e., similar in spirit to Principal Component Regression).
  But: sparse, interpretable descriptors
- Feature selection.
  But: non-greedy solver
- Symbolic regression
  But: deterministic solver

It is also a basis-set selection method
Compressed sensing: the quest for descriptors and predictive models

Structure map with SISSO, starting from 7 atomic + 6 dimer features
Feature space: $10^{11}$ features

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

- RS, $\Delta \leq -0.2$ eV
- RS, $-0.2$ eV $< \Delta \leq -0.1$ eV
- RS, $-0.1$ eV $< \Delta \leq -0.05$ eV

$\Delta = 1.0$ eV

$|P(A)|r_s(B)(1-E_s(A)/E_s(B)) - E_A(B) - E_s(A)|$ [eVÅ]

$r_s(B) / (r_p(A) + r_p(B)) / (r_s(A)r_p(A) + r_p(B)^2)$ [Å$^{-2}$]

R. Ouyang et al. PRM (2018)
Cross validation

- Iterated random selection of a subset of the data for training and test on the left out set
- Typically used for tuning "hyperparameters"
Cross validation

- Iterated random selection of a subset of the data for training and test on the left out set
- Typically used for tuning “hyperparameters”
- In compressed sensing the hyperparameters are
  - the level of sparsity (optimal dimensionality of the model)
  - the size of the feature space, and
  - (specific to SISSO) the size of the selected subsets
Cross validation

- Iterated random selection of a subset of the data for training and test on the left out set
- Typically used for tuning “hyperparameters”
- In compressed sensing the hyperparameters are
  - the level of sparsity (optimal dimensionality of the model)
  - the size of the feature space, and
  - (specific to SISSO) the size of the selected subsets

![Graph showing CV RMSE (meV/atom) vs. descriptor dimension]

R. Ouyang et al. PRM (2018)
- Iterated random selection of a subset of the data for training and test on the left out set.
- Typically used for tuning “hyperparameters”.
- In compressed sensing the hyperparameters are:
  - the level of sparsity (optimal dimensionality of the model),
  - the size of the feature space, and
  - (specific to SISSO) the size of the selected subsets.

The optimal dimensionality:
- mildly depends on the size of the selected subset
- decreases with increasing size of the feature space.
New cost function to be minimized: overlap of convex domains

1. # points in the convex overlap domain
2. Area of the domain overlap
3. Distance between domains

Good also for multi-categorical problems (see A. F. Bialon et al., Chem. Mater. 28, 2550 (2016))
SISSO: metal/nonmetal classification of binary materials

**Challenge:**
Given the formula \( A_xB_y \) of a binary material AND its crystal structure, is it a metal or a nonmetal?

**Dataset:**
~300 materials from *Springer Materials*
- \( B \) is a \( p \)-block element, \( A \) any element
- 3D materials (i.e., not layered)
- At least one 1\(^{st} \) neighbor of \( A(B) \) is \( B(A) \)
  (i.e., no materials containing “clusters” of \( A \) and/or \( B \))

**Classification AND primary features from experiments:**
ionization energy, electron affinity, 
(Pauling) electronegativity, 
covalent radius, 
valence, atomic fraction, 
\( AB \) interatomic distance, 
cell volume normalized by the sum of atomic volumes
We have a dream

PROOF OF CONCEPT: Descriptor for the Classification “Zincblende/Wurtzite or Rocksalt?”

R. Ouyang et al. PRM (2018)

SISSO: metal/nonmetal classification of binary materials

$\chi_A^2 \mid 1 - 2x_A \chi_B / x_A$

$V_{cell}$

$x$ Atomic fraction

$IE$ Ionization energy

$\chi$ Electronegativity

$\sum V_{atom} / V_{cell} \chi_A$, $IE_B \sqrt{\chi_B} / \chi_A$ (eV)

R. Ouyang et al. PRM (2018)
We have a dream

**Proof of Concept:**

Descriptor for the Classification

"Zincblende/Wurtzite or Rocksalt?"

- **HgTe** (std pressure, ZB)
- **GaAs** (std pressure, ZB)
- **CdTe** (std pressure, ZB)

**SISSO:** metal/nonmetal classification of binary materials

R. Ouyang et al. PRM (2018)
SISSO: metal/nonmetal classification of binary materials

\[ \frac{\sum V_{\text{atom}} / V_{\text{cell}}}{E_B \sqrt{\chi_B}} \]

R. Ouyang et al. PRM (2018)
Perovskites’ stability: an improved Goldschmidt Tolerance Factor

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

Goldschmidt* stable perovskites: \( 0.825 < t < 1.059 \), accuracy 79%
Perovskites’ stability: an improved Goldschmidt Tolerance Factor

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

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

Goldschmidt* stable perovskites: \( 0.825 < t < 1.059 \), accuracy 79%

Our stable perovskites: \( \tau < 4.18 \), accuracy 92%

We have a dream

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

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

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

1 / \mu = Octahedral factor

ABX₃

Goldschmidt* stable perovskites: 0.825 < \( t \) < 1.059, accuracy 79%

Our stable perovskites: \( \tau < 4.18 \), accuracy 92%
\( \tau < 3.31 \) or \( \tau > 5.92 \), 99% accuracy (1/3 of the training data)
\( \tau < 3.31 \) or \( \tau > 12.08 \), 100% accuracy (1/4 of the training data)

Improved Goldschmidt Tolerance Factor: Materials design

\[ t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)} \quad \Rightarrow \quad \tau = \frac{r_X}{r_B} - n_A \left( n_A - \frac{r_A / r_B}{\ln(r_A / r_B)} \right) \]

\[ \rho \]

\[ n_A \]

\[ X = O^{2-} \]

\[ n_A = 1^+ \]

\[ X \]
We have a dream

Proof of Concept:

"Zincblende/Wurtzite or Rocksalt?"

Improved Goldschmidt Tolerance Factor:

Extension of the materials space

\[ t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)} \quad \rightarrow \quad \tau = \frac{r_X}{r_B} - n_A \left( n_A - \frac{r_A/r_B}{\ln(r_A/r_B)} \right) \]

\( (\text{CH}_3\text{NH}_3)_2BB'\text{Br}_6 \)

\( \text{Cs}_2BB'\text{Cl}_6 \)
Methods featured in this lecture:

**Compressed sensing** for building low-dimensional maps

**Intermezzo: crow-sourcing data-driven materials science**
*The NOMAD-kaggle competition*

**Subgroup discovery** for finding outstanding subsets of data

**Convolutional neural networks** for structure (image) recognition

Applications:
- prediction of ground-state (and metastable) crystal structure(s) from the composition
- predicting whether a (binary) material is a metal or insulator
- predicting novel perovskite materials
- predicting properties of transparent conducting oxides
- classifying pristine and defected atomic structures
Is domain knowledge necessary?
nomad-coe.eu/outreach/nomad-kaggle-2018

Research Prediction Competition

Nomad2018 Predicting Transparent Conductors
Predict the key properties of novel transparent semiconductors

883 teams · 24 days ago

Overview Data Kernels Discussion Leaderboard Rules

Late Submission
• Materials class: $\text{Al}_x\text{Ga}_y\text{In}_z\text{O}_3$, $x + y + z = 3$, in 6 crystal prototypes, up to 96 atoms per unit cell
We have a dream

**Proof of Concept:**

Descriptor for the Classification

"Zincblende/Wurtzite or Rocksalt?"

- Materials class: \(\text{Al}_x\text{Ga}_y\text{In}_z\text{O}_3\), \(x + y + z = 3\), in 6 crystal prototypes, up to 96 atoms per unit cell

---

Is domain knowledge necessary?

nomad-coe.eu/outreach/nomad-kaggle-2018
• Materials class: $\text{Al}_x \text{Ga}_y \text{In}_z \text{O}_3$, $x + y + z = 3$, in 6 crystal prototypes, up to 96 atoms per unit cell

• Properties: Formation energy and band gap

• Input: lattice vectors and coordinates*

• ~3000 data points (2400 training + 600 validation set) calculated with PBE, via FHI-aims
Is domain knowledge necessary?
nomad-coe.eu/outreach/nomad-kaggle-2018
Is domain knowledge necessary?
nomad-coe.eu/outreach/nomad-kaggle-2018

1st  **Takenori Yamamoto (Tony Y.)**
Descriptor: Crystal-graph \(n\)-grams
ML method: Kernel ridge regression

2nd  **Dr. Yury Lysogorskiy**
Descriptor: 7000 average and local-order scalar features
ML method: Gradient boosting regression trees

3rd  **Lars Blumenthal**
Descriptor: based on (G. Csanyi et al.’s) SOAP
ML method: Neural network
Is domain knowledge necessary?

nomad-coe.eu/outreach/nomad-kaggle-2018

GaIn$_3$O$_6$
1\textsuperscript{st} \n-grams + KRR

2\textsuperscript{nd} Scalar local-order + GBRT

3\textsuperscript{rd} SOAP* + aNN
1\textsuperscript{st} 
\textit{n-grams + KRR}

2\textsuperscript{nd} 
Scalar local-order + GBRT

3\textsuperscript{rd} 
SOAP* + aNN
Methods featured in this lecture:

**Compressed sensing** for building low-dimensional maps

*Intermezzo: crow-sourcing data-driven materials science*

*The NOMAD-kaggle competition*

**Subgroup discovery** for finding outstanding subsets of data

**Convolutional neural networks** for structure (image) recognition

Applications:
- prediction of ground-state (and metastable) crystal structure(s) from the composition
- predicting whether a (binary) material is a metal or insulator
- predicting novel perovskite materials
- predicting properties of transparent conducting oxides
- classifying pristine and defected atomic structures
Subgroup discovery: finding descriptive statements about outstanding groups

\[ y(i) = 0.52x(i) + 0.13 \]
Subgroup discovery: finding descriptive statements about outstanding groups
Subgroup discovery: finding descriptive statements about outstanding groups

We have a dream

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

Subgroup discovery: finding descriptive statements about outstanding groups
Subgroup discovery: finding descriptive statements about outstanding groups

**Ingredients:**
- Population $P = \{1, \ldots, n\}$
- Target Variable $y: P \rightarrow Y$
- Description variables $x_j: P \rightarrow X_j$
- Basic propositions $\Pi = \{\pi_1, \ldots, \pi_k\}$
- Objective functions: $f$
- {All possible subgroups of $P} \rightarrow \mathbb{R}$
Subgroup discovery: finding descriptive statements about outstanding groups

**Ingredients:**
Population \( P = \{1, \ldots, n\} \)
Target Variable \( y: P \rightarrow Y \)
Description variables \( x_j: P \rightarrow X_j \)
Basic propositions \( \Pi = \{\pi_1, \ldots, \pi_k\} \)
Objective functions: \( f \)
\{All possible subgroups of \( P\} \rightarrow \mathbb{R} \)

**Task:**
Finding \( \sigma(i) = \pi_1(i) \land \ldots \land \pi_m(i) \)
For which \( f(P) = \max \)

Typical form of \( f \):
“Size of subgroup” \( \times \) “Reduction of variance of \( Y \) compared to the whole population”
Subgroup discovery in practice

Distribution of adsorption energies of CO$_2$ on different surfaces of several metal-oxides
Subgroup discovery in practice

Distribution of adsorption energies of CO$_2$ on different surfaces of several metal-oxides
Subgroup discovery in practice

\[
\arg\max_{SG \subseteq P} U = \frac{\#SG}{\#P} \left(1 - \frac{\text{mad}(SG)}{\text{mad}(P)}\right)|\text{med}(SG) - \text{med}(P)|
\]
Subgroup discovery in practice

\[
\arg\max_{SG \subseteq P} U = \frac{\#SG}{\#P} \left(1 - \frac{\text{mad}(SG)}{\text{mad}(P)}\right) |\text{med}(SG) - \text{med}(P)|
\]
Subgroup discovery in practice

\[
\text{argmax}_{SG \subseteq P} U = \frac{\#SG}{\#P} \left( 1 - \frac{\text{mad}(SG)}{\text{mad}(P)} \right) |\text{med}(SG) - \text{med}(P)|
\]

Size of subgroup \(SG\)

Size of full set \(P\)

Mean absolute deviation from the median (spread of distribution)
Subgroup discovery in practice

\[
\text{argmax}_{SG \subset P} \quad U = \frac{\#SG}{\#P} \left( 1 - \frac{\text{mad}(SG)}{\text{mad}(P)} \right) |\text{med}(SG) - \text{med}(P)|
\]

Size of subgroup \(SG\)

Mean absolute deviation from the median (spread of distribution)

Size of full set \(P\)

Median of the distribution

Subgroup discovery in practice

\[ \arg\max_{SG \subseteq P} U = \frac{\#SG}{\#P} \left( 1 - \frac{\text{mad}(SG)}{\text{mad}(P)} \right) \]

- **Mean absolute deviation from the median** (spread of distribution)
- **Maximize median shift**
- **Median of the distribution**
- **Minimize relative spread of \( SG \)**


Subgroup discovery in practice

The size of subgroup $SG$ is described by a selector, a conjunction of statements ($s_1 \land s_2 \land \ldots$) about a list of given features e.g.,

- $s_1 = \text{surface energy larger than} \ldots$
- $s_2 = \text{p-band center of surface O less than} \ldots$

The objective is to maximize the median shift and minimize the relative spread of $SG$.

$$\arg\max_{SG \subseteq P} U = \frac{\#SG}{\#P} \left( 1 - \frac{\text{mad}(SG)}{\text{mad}(P)} \right)$$

- Mean absolute deviation from the median (spread of distribution)
- Maximize median shift
- Minimize relative spread of $SG$
Subgroup discovery in practice

\[ \arg\max_{SG \subset P} U = \frac{\#SG}{\#P} \left( 1 - \frac{\text{mad}(SG)}{\text{mad}(P)} \right) \left| \text{med}(SG) - \text{med}(P) \right| \]

- CO\text{\textsubscript{2}} adsorption energy [eV]
- Data points / bin
Subgroup discovery in practice

\[
\arg\max_{SG \subseteq P} \ U = \frac{\#SG}{\#P} \left( 1 - \frac{\text{mad}(SG)}{\text{mad}(P)} \right) \left| \text{med}(SG) - \text{med}(P) \right|
\]
Subgroup discovery in practice

The (SISSO) model for the discovered subgroup
- is more accurate than the global model
- has a different descriptor due to different physics.

Small work function:
Surfaces with dominantly ionic character
Methods featured in this lecture:

**Compressed sensing** for building low-dimensional maps

*Intermezzo: crow-sourcing data-driven materials science*

*The NOMAD-kaggle competition*

**Subgroup discovery** for finding outstanding subsets of data

**Convolutional neural networks** for structure (image) recognition

Applications:
- prediction of ground-state (and metastable) crystal structure(s) from the composition
- predicting whether a (binary) material is a metal or insulator
- predicting novel perovskite materials
- predicting properties of transparent conducting oxides
- classifying pristine and defected atomic structures
Automatic recognition of (crystal) structures

- Given an atomic arrangement, determine the ("most similar") prototype among the following classes:
The motivations

- Automatic **nanomaterial classification** of the NOMAD Archive without user-defined thresholds and robust to defects

<table>
<thead>
<tr>
<th></th>
<th>Total-Energy Calculations</th>
<th>Bulk Crystals</th>
<th>Surfaces</th>
<th>Molecules/Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>44,179,006</td>
<td>39,402,326</td>
<td>220,918</td>
<td>4,517,016</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Different Geometries</th>
<th>Chemical Compositions</th>
<th>Band Structures</th>
<th>Phonon Calculations</th>
</tr>
</thead>
<tbody>
<tr>
<td>40,481,615</td>
<td>281,135</td>
<td>1,936,325</td>
<td>91</td>
<td></td>
</tr>
</tbody>
</table>

- Automatic **recognition of nano-islands** in *atom probe tomography* experiments

The diffraction fingerprint: intuition

- Rotate the crystal structure of 45° and (-45°) about the x, y, and z axis

- Calculate the diffraction fingerprint for each rotation:
  - around x-axis
  - around y-axis
  - around z-axis

- Sum the results in a RGB image
The diffraction fingerprint: results

- Size-invariant (structures of all sizes are learned by construction)
- Robust to “perturbation” of the structure
- Same dimension (e.g. 64x64 pixels) regardless of the number of atoms
- Easily interpretable
- Different diffraction patterns for each class (and similar within the same class)
The diffraction fingerprint: results

- Body-centered-tetragonal (spgroup=139)
- Body-centered-tetragonal (spgroup=141)
- Rhombohedral/Hexagonal (spgroup=166/194)
- Simple cubic (spgroup=221)
- Face-centered-cubic (spgroup=225)
- Diamond (spgroup=227)
- Body-centered-cubic (spgroup=229)
The method workflow

Crystal structure (from NOMAD Archive)

representation

Image (diffraction fingerprint)

classification

Crystal class (or “most similar”)

Representation:
given a structure we calculate its diffraction fingerprint

Prediction model:
convolutional neural network (state-of-the-art for image recognition)
ConvNets: human analogy

How do we (humans) subconsciously classify an image?
Looking for identifiable (pre-learned) features (e.g. for dogs: paws, 4 legs)

How does a computer classify an image?
Looking at low level features (edges and curves), and then build more abstract concepts though a series of (convolutional) layers.

Convolutional Layer: intuition

- Example: curve detection

Pixel representation of filter

Visualization of a curve-detector filter

Convolutional Layer: intuition

Output: **high value** because the shape (mouse) resembles the filter (curve)

Ref: https://adeshpande3.github.io/A-Beginner's-Guide-To-Understanding-Convolutional-Neural-Networks/
The output of a convolutional layer is a **2D activation map** that gives the responses of that filter at every spatial position.

ConvNets in practice

- Each conv. layer has multiple filters (e.g. curve, horizontal line, etc...)
- Multiple conv. Layers:
  - 1\textsuperscript{st} layer:
    - input=image → low-level filters (e.g. curve or straight edges)
  - 2\textsuperscript{nd} layer:
    - input=activation map → higher level filters (e.g. semicircles: curve+straight edges, squares)
  - n\textsuperscript{th} layer: high level filters (e.g. paws)

- Complexity of the filters increases layer by layer
- Filters learned minimizing the training error
Our ConvNet

<table>
<thead>
<tr>
<th>Layer type</th>
<th>Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolutional Layer</td>
<td>(Kernel: 9x9; 32 filters)</td>
</tr>
<tr>
<td>Convolutional Layer</td>
<td>(Kernel: 9x9; 16 filters)</td>
</tr>
<tr>
<td>Max Pooling Layer</td>
<td>(Pool size: 2x2, stride: 2x2)</td>
</tr>
<tr>
<td>Convolutional Layer</td>
<td>(Kernel: 9x9; 8 filters)</td>
</tr>
<tr>
<td>Convolutional Layer</td>
<td>(Kernel: 9x9; 8 filters)</td>
</tr>
<tr>
<td>Max Pooling Layer</td>
<td>(Pool size: 2x2, stride: 2x2)</td>
</tr>
<tr>
<td>Convolutional Layer</td>
<td>(Kernel: 9x9; 4 filters)</td>
</tr>
<tr>
<td>Convolutional Layer</td>
<td>(Kernel: 9x9; 4 filters)</td>
</tr>
<tr>
<td>Fully connected Layer + Dropout</td>
<td>(Size: 128; dropout: 25%)</td>
</tr>
<tr>
<td>Batch Normalization</td>
<td>(Size: 128)</td>
</tr>
<tr>
<td>Softmax</td>
<td>(Size: 7)</td>
</tr>
</tbody>
</table>
The pristine dataset

- Dataset 1: AFLOWlib elemental solids
  - Includes ~90 chemical elements
  - Supercells of various sizes

- Dataset numbers:
  - 10,675 images; 7 classes
  - 90% training, 10% validation (randomly)
  - ConvNet runtime (laptop): train: ~80min, pred. ~70 ms @img

<table>
<thead>
<tr>
<th>Training accuracy [%]</th>
<th>Validation accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

The defective dataset (test set)

- **Dataset 2**: dataset 1 with added defects
  - *Random displacements*: up to st.dev. 0.08 Å
  - *Random vacancies*: up to 25%
  - *Substitutions* (randomly change the type of atom: e.g. C -> H)

- **Dataset numbers**:
  - 139,945 images
  - 7 classes

<table>
<thead>
<tr>
<th>Training accuracy [%]</th>
<th>Test accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Training</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Model exploration 1/2

Body-centered-cubic structure  Intermediate structure  Simple cubic structure

Model exploration 2/2


Back-projection to image space

- Convolution → Deconvolution
- Pooling → Unpooling

Understanding ConvNets


Understanding ConvNets

Attentive response maps (RGB)


Acknowledgements

Compressed sensing, SISSO, and metal/insulator proof of concepts
Jan Vybiral, Runhai Ouyang, Emre Ahmetcik, Stefano Curtarolo, Sergey Levchenko, Claudia Draxl

Application of SISSO to perovskites
Christopher J. Bartel, Christopher Sutton, Bryan R. Goldsmith, Runhai Ouyang, Charles B. Musgrave

Transparent conducting oxide: NOMAD-kaggle competition
Christopher Sutton, Angelo Ziletti, Claudia Draxl, Daan Frenkel, Kristian Thygesen, Samuel Kaski, Bernhard Schölkopf

Subgroup Discovery and application to CO₂ adsorption
Mario Boley, Jilles Vreeken, Aleksei Mazheika, Sergey Levchenko

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

All the above
Matthias Scheffler

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