MATERIALS PROPERTY PREDICTIONS USING MACHINE LEARNING:
RECENT EXAMPLES AND FUTURE OUTLOOK

Ghanshyam Pilania
Theory Department
Fritz Haber Institute of the Max Planck Society
&
Materials Science and Technology (MST) Division
Los Alamos National Laboratory

NOMAD SUMMER: A HANDS-ON COURSE ON TOOLS FOR NOVEL-MATERIALS DISCOVERY
28th September 2017, 9:00 AM
Physics Department (Lise-Meitner-Haus), Humboldt-University, Berlin
AGE OF DATA, ML & AI

- Transportation
- Recommender systems
- Facial recognition
- Board games
- Medical applications
- Robotics
- Speech recognition
- Finance and marketing
INFORMATICS: IN THE PAST

Steel production (India & Sri Lanka)

6th Century BC

Kepler’s Laws (Johannes Kepler)

17th Century

Data-driven Nursing (Florence Nightingale)

19th Century

Periodic Table (Dimitri Mendeleev)

19th Century

Hume-Rothery Rules Hall-Petch Relationship

Mid 20th Century

Chem/Bio Informatics Polymer Informatics

Late 20th Century

Materials Informatics

Early 21st Century
DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.

The areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent areas for annual deaths from Preventible or Notifiable diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov. 1854 marks the boundary of the deaths from all other causes during the month.
In Oct. 1854, & April 1855, the black area coincides with the red; in January & February 1855 the blue coincides with the black.
The entire areas may be compared by following the line, the red & the black lines enclosing them.

INFORMATICS: IN THE PAST

- Steel production (India & Sri Lanka)
- Kepler’s Laws (Johannes Kepler)
- Data-driven Nursing (Florence Nightingale)
- Periodic Table (Dimitri Mendeleev)
- Hume-Rothery Rules
- Hall-Petch Relationship
- Chem/Bio Informatics
- Polymer Informatics
- Materials Informatics

Timeline:
- 6th Century BC
- 17th Century
- 19th Century
- 19th Century
- Mid 20th Century
- Late 20th Century
- Early 21st Century
MATERIALS BIG-DATA
A PARADIGM SHIFT

1st paradigm: Empirical science

ΔU = Q - W
Change in heat (Q) - work (W) internal added done energy to system by system

2nd paradigm: Model-based theoretical science

Laws of Thermodynamics

Experiments

3rd paradigm: Computational science (simulations)

Density Functional Theory, Molecular Dynamics

4th paradigm: (Big) data driven science

Predictive analytics
Clustering
Relationship mining
Anomaly detection

A. Agrawal and A. Choudhary, APL Mater. 4, 053208 (2016)
MATERIALS INFORMATICS: TODAY

- Computational Power
- Algorithms & Methods Development
- High throughput Exp.
- Materials Genome Initiative

Moore’s Law

Open-Source Materials Databases

- The Materials Project
- NOMAD
- Khazana 1.0
- OpenKIM
- NIST
- AFLOW
- NREL
- OQMD
- Citrine
- Scientific Data

and many more…
DATA TO INSIGHTS AND PREDICTIONS

- Efficient enumeration
- Targeted Search
- Adaptive Optimization

Design

Materials

Laborious computations and/or Experiments

Fingerprint

Property

Machine Learning

RECENT HIGHLIGHTS

Designing Polymer Dielectrics for Energy Storage

Learning Bandgaps Solids

Learning Models for Dielectric Breakdown Strength

Understanding Radiation Damage Resistance


Chem. Mater. 29, 2574 (2017)
• Phenomenological model discovery for **intrinsic dielectric breakdown strength** of insulators using machine learning

![Diagram](image)


• Multi-fidelity machine learning models for **bandgap** prediction

![Diagram](image)

• Phenomenological model discovery for **intrinsic dielectric breakdown strength** of insulators using machine learning

**EXAMPLE 1**

DIELECTRIC BREAKDOWN

Rapid reduction of resistance of an electrical insulator under the presence of extreme electric field
Predicting intrinsic electrical breakdown field of an insulator from first principles is difficult…

Determined by the balance between energy gain (E-field to e) and loss (e to phonon) of the electron

Can the breakdown field be estimated rapidly using a simple heuristic model?

Consider 82 binary octets (ex. ZnO, NaCl, …)

Fröhlich, Nature 151, 339 (1943)
FIRST PRINCIPLES CALCULATIONS

Alkali metal halides
Post-transition metal halides
Transition metal halides
Alkaline earth metal chalcogenides

Transition metal oxides
Group IV
Group III-V
Group II-VI

LEARNING FROM DATA

Intrinsic breakdown field of 82 binary octets (by DFT)

Easily accessible material properties

- Band gap ($E_g$)
- Phonon cutoff frequency ($\omega_{\text{max}}$)
- Mean phonon frequency ($\omega_{\text{mean}}$)
- Bulk modulus ($M$)
- Dielectric constant, electron ($\varepsilon_e$)
- Dielectric constant, total ($\varepsilon_{\text{tot}}$)
- Nearest neighbor distance ($d_{\text{NN}}$)
- Density ($\rho$)

Correlation analysis & Machine learning

$F_b = f(A, B, \ldots)$
FEATURE CREATION

8 primary features
- Band gap ($E_g$)
- Phonon cutoff frequency ($\omega_{\text{max}}$)
- Mean phonon frequency ($\omega_{\text{mean}}$)
- Bulk modulus ($M$)
- Dielectric constant, electron ($\varepsilon_e$)
- Dielectric constant, total ($\varepsilon_{\text{tot}}$)
- Nearest neighbor distance ($d_{\text{NN}}$)
- Density ($\rho$)

12 prototypical functions
- $x, x^1, x^2, x^3, x^2, x^{-1/2}$
- $\ln(x), 1/\ln(x), \exp(x), x^{-x}$

96 features with 1 function
- ex: $\ln(E_g)$

4,480 unique features with 2 functions
- ex: $\varepsilon_{\text{tot}}^2/\exp(d_{\text{NN}})$

183,368 unique features with 3 functions
- ex: $\ln(\omega_{\text{max}})\exp(M)/E_g^2$

187,952 compound features

FEATURE SELECTION USING **LASSO**

187,952 compound features

**LASSO-based down-selection**

1. **Feature**\(_n\) (n=1~187,952)
2. Highly correlated? (based on **LASSO** coefficient)
   - Yes → Survive
   - No → Discard

**36 features**

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Feature</th>
<th>Absolute Pearson correlation (\frac{\omega}{\ln F_b})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(\ln E_g \ln \omega_{\max} / d_{NN}^{1/2})</td>
<td>0.899</td>
</tr>
<tr>
<td>2</td>
<td>((E_g \omega_{\max})^{1/2})</td>
<td>0.890</td>
</tr>
<tr>
<td>3</td>
<td>(\ln E_g \omega_{\max}^{1/2})</td>
<td>0.890</td>
</tr>
<tr>
<td>4</td>
<td>(E_g^{1/2} \ln \omega_{\max})</td>
<td>0.889</td>
</tr>
<tr>
<td>5</td>
<td>(E_g^{1/2} / d_{NN})</td>
<td>0.885</td>
</tr>
<tr>
<td>6</td>
<td>(\ln E_g / d_{NN}^2)</td>
<td>0.883</td>
</tr>
<tr>
<td>7</td>
<td>(\ln E_g / \exp(d_{NN}))</td>
<td>0.880</td>
</tr>
<tr>
<td>8</td>
<td>(E_g^{1/2} / \ln d_{NN})</td>
<td>0.879</td>
</tr>
<tr>
<td>9</td>
<td>(\omega_{\max} E_g^{1/2} / \ln \omega_{\text{mean}})</td>
<td>0.871</td>
</tr>
<tr>
<td>10</td>
<td>((\omega_{\max} / E_g)^{1/2})</td>
<td>0.869</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>36</td>
<td>((\varepsilon_{\text{tot}} E_g)^{1/2})</td>
<td>0.480</td>
</tr>
</tbody>
</table>

* **LASSO**: Least absolute shrinkage and selection operator
PREDICTION MODEL

Model formula:
\[ F_b = 24.442 \exp\{0.315(E_g \omega_{\text{max}})^{1/2}\} \]

Predicted intrinsic breakdown field in MV/m
Band gap in eV
Phonon cutoff frequency in THz

Entirely based on heuristic not a law!
But practically useful in estimating electrical breakdown field strength.

APPLICATION TO PEROVSKITES

209 Perovskites

Prediction of breakdown field

Compounds with highest breakdown field

Boron containing compounds appear highly promising

SUMMARY: EXAMPLE 1

Intrinsic dielectric breakdown field of 82 binary octets are obtained by using quantum mechanical calculations.

Phenomenological models as a function of $E_g$ & $\omega_{\text{max}}$ are developed using machine learning method.

Application to perovskites predicts boron containing compounds as promising high breakdown strength materials.
• Multi-fidelity machine learning models for **bandgap** prediction

MULTI-FIDELITY INFORMATION FUSION

Given limited computational resources, how to tune the cost-accuracy trade-off for optimal predictions?

Implications for materials genomics:

➢ High through chemical space explorations
➢ Rational materials design and discovery
**MULTI-FIDELITY LEARNING FOR BANDGAPS OF SOLIDS**

- A property of interest for many applications, including energy harvesting, energy storage, catalysis, scintillation and device physics
- A natural hierarchy of “DFT and beyond” approaches provides different options for the “cost-accuracy” trade-offs

  ➢ Standard local and semi-local XC functionals do not provide a good description of the bandgaps
  ➢ Hybrid functionals and beyond-DFT techniques are extremely expensive

---


---

Jacob's ladder of DFT exchange-correlation (XC) functionals by John P. Perdew
AN EXAMPLE OF ELPASOLITES

- A class of Materials with potential applications in energy harvesting and scintillation
- Exhibiting flexible chemistry, amenable to combinatorial synthesis
- Variable chemistry on a fixed cubic lattice makes this class an ideal test case for machine learning
- A dataset of 600 Elpasolites
- DFT computed PBE bandgaps used as low fidelity data and HSE06 bandgaps were taken as high fidelity data

<table>
<thead>
<tr>
<th>Computed</th>
<th>Training</th>
<th>Validation</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>600 PBE</td>
<td>200 PBE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>250 HSE</td>
<td>200 HSE</td>
<td>50 HSE</td>
<td>350 HSE</td>
</tr>
</tbody>
</table>

The Feature Space

- Elemental electronegativity,
- First ionization potential,
- Empirical radius and
- Pettifor's Mendeleev number
(for each of the species occurring at A, B, B' and X sites)

DETAILS OF THE LEARNING MODEL

- We employ a co-kriging model within a Bayesian framework

\[
Z_c(x) = \rho Z_c(x) + Z_d(x) \quad \exists(x, x', \theta) = \exp \left\{ -\sum_{j=1}^{d} \theta_j \|x_j - x'_j\|^2 \right\}
\]

Independent Gaussian Processes

\[
K =
\begin{pmatrix}
\sigma_c^2 \Theta_c(X_c, X_c) & \rho \sigma_c^2 \Theta_c(X_c, X_e) \\
\rho \sigma_c^2 \Theta_c(X_c, X_e) & \rho^2 \sigma_c^2 \Theta_c(X_e, X_e) + \sigma_d^2 \Theta_d(X_e, X_e)
\end{pmatrix}
\]

\[
\mu_e^* = \hat{\mu} + k^T K^{-1}(y - 1\hat{\mu}) \\
\sigma_e^2 = \rho^2 \hat{\sigma}_c^2 + \hat{\sigma}_d^2 - k^T K^{-1}k,
\]

\[
\hat{\mu} = 1^T K^{-1} y / 1^T K^{-1} 1 \\
k = \begin{pmatrix}
\rho \sigma_c^2 \Theta_c(X_c, x^*) \\
\rho^2 \hat{\sigma}_c^2 \Theta_c(X_e, x^*) + \hat{\sigma}_d^2 \Theta_d(X_e, x^*)
\end{pmatrix}
\]

THE COST-ACCURACY TRADE-OFF

RMS Error (eV) on the Validation Set (Unseen Data)

<table>
<thead>
<tr>
<th>Training</th>
<th>RMS Error (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 PBE</td>
<td>180 HSE</td>
</tr>
<tr>
<td>200 HSE</td>
<td>110 HSE</td>
</tr>
<tr>
<td></td>
<td>130 PBE</td>
</tr>
</tbody>
</table>

THE COST-ACCURACY TRADE-OFF

CURRENT STATE-OF-THE-ART

**Linear fit DFT v/s experiments**


**Use PBE bandgap as a feature in ML**

Lee et al. Physical Review B 93, 115104 (2016).


A reliable Intermediate Filter for Application Specific Screening

Results are averaged over the B-site

PRACTICALLY USEFUL FOR CHEMICAL SPACE EXPLORATIONS

NEXT CRITICAL STEPS...

- Use uncertainties for adaptive design (active learning)
- Tune the trade-offs between exploration and exploitation
NEXT CRITICAL STEPS...

• Use uncertainties for adaptive design (active learning)

• Tune the trade-offs between exploration and exploitation

• Multi-objective optimization to systematically explore Pareto-optimal set

• Improved ways of incorporating domain knowledge in machine learning models
NEXT CRITICAL STEPS . . .

LEARNING FORM OTHER COMMUNITIES

Materials problems are different!
Still, many already existing methods can be useful to solve materials challenges
Acknowledgements

Co-workers & Collaborators:

J. Gubernatis  T. Lookman  J. Theiler  A. K. M.-Kanakkithodi  C. Kim  R. Ramprasad

Hosts at the Fritz Haber Institute

M. Scheffler  L. Ghiringhelli

Funding & Computational Resources

Los Alamos National Laboratory
Alexander von Humboldt Foundation
Discover LDRD
XSEDE