MACHINE LEARNING IN MATERIALS SCIENCE

Rampi Ramprasad
University of Connecticut

http://rampi.ims.uconn.edu
WHAT IS MACHINE LEARNING?

• **The context:** A pattern exists ... but we cannot pin it down mathematically ... we have data on it though ...

• **The concept:** Build a heuristic, predictive model ... based purely on past experience/data ... rather than explicitly solving equations underlying known laws

• The learning process should be “adaptive”, i.e., as more data accumulates, the predictive power should increase
EXAMPLE: MOVIE RATING

Courtesy: Learning from data, Abu-Mustafa, Magnon-Ismail & Lin
EXAMPLE: CREDIT APPROVAL

Applicant information

<table>
<thead>
<tr>
<th>age</th>
<th>23 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>male</td>
</tr>
<tr>
<td>annual salary</td>
<td>$30,000</td>
</tr>
<tr>
<td>years in residence</td>
<td>1 year</td>
</tr>
<tr>
<td>years in job</td>
<td>1 year</td>
</tr>
<tr>
<td>current debt</td>
<td>$15,000</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Approve credit?

Courtesy: Learning from data, Abu-Mustafa, Magnon-Ismail & Lin
**LEARNING COMPONENTS**

Input: \( x \)  
(customer application)

Output: \( y \)  
(good/bad customer?)

Data: \((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\)  
(historical records)

Hypothesis: \( g(x) = y \)  
(To be determined, refined, and used in the future)

“Training set” versus “test set”
WHEN TO USE MACHINE LEARNING?

• When fundamental laws underlying a process don't exist [e.g., social science problems]

• When such fundamental laws may exist, but are enormously complex [e.g., weather prediction]

• When we have a lot of data and we are looking for simple rules and correlations [e.g., Hall-Petch equation]
“Machine learning in materials science: Recent progress and emerging applications,”
Mueller, Kusne & Ramprasad, Reviews of Computational Chemistry (2016)

“Atomistic calculations and materials informatics: A review”
FINGERPRINT IS KEY!

• The “fingerprint” is a numerical representation of the material

• It should be defined based on the application & domain knowledge

• It should be intuitive, and inexpensive to compute

• It should be invariant to transformations of the material, such as translation, rotation, and permutations of like elements

• Fingerprints can be macroscopic or microscopic
EXAMPLE 1: COARSE FINGERPRINTS
PREDICTING ELECTRICAL BREAKDOWN

Predicting the *intrinsic* electrical breakdown field of an insulator from first principles is difficult …

… it is determined by the balance between energy gained by an electron from the field and energy lost to phonons …

… but can the breakdown field be estimated rapidly using a simple heuristic model?

*Dependence on Chemistry?*

… consider 82 binary octets

Frohlich, *Nature* 151, 339 (1943)
LEARNING FROM DATA

Intrinsic breakdown field of 82 binary octets

Easily accessible material properties

Band gap
Phonon cutoff frequency
Mean phonon frequency
Dielectric constant (electronic)
Dielectric constant (total)
Nearest neighbor distance
Density
Bulk modulus

$F_b = f(A, B, ...)$?

Correlation analysis & Machine learning
8 Primary features

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

12 prototype functions

- $x$
- $1/x$
- $x^2$
- $x^3$
- $x^{3/2}$
- $\sqrt{x}$
- $1/\sqrt{x}$
- $\ln(x)$
- $1/\ln(x)$
- $e^x$
- $e^{-x}$

96 features with 1 function

- 8 $e^x \ln(x)$
- 12 $e^x$

4,480 compound features with 2 functions

- 96 $e^x x^2/e^x$

183,368 compound features with 3 functions

- 96 $e^x x^3 \ln(x)/\sqrt{x}$

Total 187,952 features

FEATURE SELECTION WITH "LASSO"

**Total 187,952 features**

LASSO-based down-selection

Feature$_n$ (n=1~187,952)

Highly correlated? (based on LASSO coefficient)

Yes → Survive

No → Discard

**36 compound features**

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Compound Feature</th>
<th>Absolute Pearson correlation /w $\ln F_b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\ln E_g \ln \omega_{max} / \sqrt{d_{NN}}$</td>
<td>0.899</td>
</tr>
<tr>
<td>2</td>
<td>$\sqrt[4]{\omega_{max}} \sqrt{E_g}$</td>
<td>0.890</td>
</tr>
<tr>
<td>3</td>
<td>$\sqrt[4]{\omega_{max}} \ln E_g$</td>
<td>0.890</td>
</tr>
<tr>
<td>4</td>
<td>$\sqrt{E_g} \ln \omega_{max}$</td>
<td>0.889</td>
</tr>
<tr>
<td>5</td>
<td>$\sqrt{E_g} / d_{NN}$</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>$\sqrt{E_g} / \ln d_{NN}$</td>
<td>0.879</td>
</tr>
<tr>
<td>9</td>
<td>$\omega_{max} \sqrt{E_g} / \ln \omega_{mean}$</td>
<td>0.871</td>
</tr>
<tr>
<td>10</td>
<td>$\omega_{max} / \sqrt{E_g}$</td>
<td>0.869</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>36</td>
<td>$\sqrt{e_{tot}} \sqrt{E_g}$</td>
<td>0.480</td>
</tr>
</tbody>
</table>
PREDICTION (& DESIGN?)

The finding

\[ F_b = 24.442 \exp(0.315 \sqrt{E_g \omega_{\text{max}}}) \]

Note: This is a heuristic, not a law; we cannot take it too seriously!
AN APPLICATION

EXAMPLE 2: MEDIUM FINGERPRINTS

PREDICTING POLYMER PROPERTIES

Organic Blocks: CH$_2$, CO, CS, O, NH, C$_6$H$_4$, C$_4$H$_2$S

Fingerprint Choice 1: “Singles” - number of CH$_2$, NH, CO, C$_6$H$_4$, etc.
Fingerprint Choice 2: “Doubles” - number of CH$_2$-CH$_2$, NH-CH$_2$, CO-NH, etc.
Fingerprint Choice 3: “Triples” - number of CH$_2$-CH$_2$-CH$_2$, NH-CH$_2$-CO, etc.

Kernel ridge regression

Measure of similarity: Euclidean distance

\[ d(i, j) = \sqrt{\left( |x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2 \right)} \]

Property Estimation:
Sum of weighted Gaussians

\[ Y^{est}(P) = \sum_{i=1}^{N} \alpha_i \exp \left[ -\frac{1}{2\sigma^2} d(P, P_i)^2 \right] \]

INSTANT PROPERTY PREDICTIONS

... using a DFT dataset for about 300 organic polymers

Mannodi-Kanakkithodi et al., *Scientific Reports* 6, 20952 (2016)
Polymer Dielectrics Property Prediction

Obtain properties predicted by DFT and/or machine learning [1,2]

Choose the building blocks (co-polymers) of polymer repeat unit.

![Building blocks interface]

Predict Properties

References


Target polymer summary

Repeat unit: -(NH-CO-NH-C6H4)_\textit{r}^{-1}
Number of building blocks (co-polymers): 4
Chemical formula: C_7H_6ON_2

Predicted properties using machine learning model

<table>
<thead>
<tr>
<th>Method</th>
<th>Polymer</th>
<th>Dielectric Constant</th>
<th>Refractive Index</th>
<th>Band Gap, HSE06 (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Learning [1]</td>
<td>C_7H_6ON_2</td>
<td>4.95</td>
<td>1.98</td>
<td>3.30</td>
</tr>
<tr>
<td></td>
<td>NH-CO-NH-C6H4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NH-CO-NH-C6H4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFT (ID: 0299)</td>
<td>O_2C_13H_2N_4</td>
<td>4.91</td>
<td>1.97</td>
<td>3.45</td>
</tr>
<tr>
<td></td>
<td>NH-CO-NH-C6H4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Polymer Dielectrics Property Prediction

Obtain properties predicted by DFT and/or machine learning [1,2]

Choose the building blocks (co-polymers) of polymer repeat unit.

Predict Properties

References


Target polymer summary

Repeat unit: \((-\text{NH-}\text{CO-}\text{NH-}\text{C}_4\text{H}_4\text{-CH}_2\text{-C}_4\text{H}_2\text{S-}\text{CS-O-}\text{CH}_2\text{-})_n\)

Number of building blocks (co-polymers): 9

Chemical formula: \(\text{C}_{14}\text{H}_{12}\text{O}_2\text{N}_2\text{S}_3\)

Predicted properties using machine learning model

<table>
<thead>
<tr>
<th>Method</th>
<th>Polymer</th>
<th>Dielectric Constant</th>
<th>Refractive Index</th>
<th>Band Gap, HSE06 (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Learning [1]</td>
<td>(\text{C}<em>{14}\text{H}</em>{12}\text{O}_2\text{N}_2\text{S}_3) (\text{NH-}\text{CO-}\text{NH-}\text{C}_4\text{H}_4\text{-CH}_2\text{-C}_4\text{H}_2\text{S-}\text{CS-O-}\text{CH}_2\text{-})</td>
<td>4.51</td>
<td>1.92</td>
<td>3.92</td>
</tr>
<tr>
<td>Machine Learning [2]</td>
<td>(\text{C}<em>{14}\text{H}</em>{12}\text{O}_2\text{N}_2\text{S}_3) (\text{NH-}\text{CO-}\text{NH-}\text{C}_4\text{H}_4\text{-CH}_2\text{-C}_4\text{H}_2\text{S-}\text{CS-O-}\text{CH}_2\text{-})</td>
<td>4.46</td>
<td>1.92</td>
<td>3.35</td>
</tr>
</tbody>
</table>
• Given only the atomic configuration, can we directly predict the atomic forces rapidly & accurately?

• Atoms respond to forces, and force is a local quantity (unlike energy)
THE CONCEPT

Potential energy
- global property of entire system

Atomic force
- local environment-dependent property

"Fingerprint"

Machine Learning

Force Field

THE ML FORCE FIELD
(AGNI)
GEOMETRY OPTIMIZATION

~ 850 Al atoms in FCC phase with vacancies
LARGE-SCALE MOTIONS

Adatom migration on Al (111)

Is the physics underlying dynamics preserved?
FURTHER VALIDATION

... on situations not used during “training”
ENERGY FROM FORCES

MD simulations of bulk Al
## Hierarchy of Fingerprints

<table>
<thead>
<tr>
<th>Scale</th>
<th>Illustrative Examples</th>
<th>Accuracy</th>
</tr>
</thead>
</table>
CRITICAL NEXT STEPS

• Decision on whether to use machine learning
• Choice of fingerprints
• Uncertainty (out-of-domain) quantification
• Automatic & adaptive improvement of model
• Show the value!
CREDITS

2013

Dr. Ghanshyam Pilania (LANL)
Dr. Chenchen Wang (UCSB/FHI)
Dr. Venkatesh Botu (Corning)

2015

Dr. Chiho Kim
Dr. Huan Tran
James Chapman
Arun M-K

DEPARTMENT OF THE NAVY
Science & Technology
Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain ...