Text and Data Mining for Material Synthesis

Elsa Olivetti, MIT  Gerbrand Ceder, UC Berkeley  
Departments of Materials Science & Engineering

Andrew McCallum, UMass Amherst  
Department of Computer Science & Engineering
Challenges for technology development:
Timeline for development is long

1954
- 1st practical silicon solar cell invented at Bell Labs

1963
- Sharp produces 1st practical solar module of silicon solar cells

1982
- Kyocera 1st mass produces polysilicon cells by today's standard process

1980
- Oxford demonstrates 1st viable rechargeable lithium battery

1991
- Sony sells 1st commercially available Li-ion batteries for high price consumer electronics

2008
- 1st uses of Li-ion battery in production vehicles
Modern data-driven and first-principles materials design accelerates pace of what to make...

\[ H = \sum_{i=1}^{N_e} \nabla_i^2 + \sum_{i=1}^{N_e} V_{nuclear}(r_i) + \frac{1}{2} \sum_{i}^{N_e} \sum_{j \neq i}^{N_e} \frac{1}{|r_j - r_i|} \]

- Phase diagrams
- Bandgaps
- Surfaces
  - 3.2 V (101)
  - 3.46 V (011)
  - 3.86 V (011)
  - 4.09 V (010)
  - 3.7 V (101)
  - 3.76 V (010)
  - 3.2 V (010)
Materials Synthesis Articles → Natural Language Processing → Compiled Synthesis Routes → Machine Learning → Recommended Synthesis Conditions

Over one million scientific articles
Text extraction workflow

PDF | HTML | XML | …… | Web-Accessible Journal Articles

Article Retrieval API
Plaintext Conversion
Paragraph and Token Classification
Synthesis Info Extraction

JSON Data Object

Metadata + Text-mined Data

Metadata Records
Synthesis Parameters
Material Properties

Synthesis + Property Database
Status of data dissemination

**Paragraph API**  **OPERATIONAL**
“Barium titanate was synthesized using a solid-state processing route. Precursors were obtained from...”

**Recipe API**  **IN-PROGRESS**
(1) Obtain precursors
(2) Mix barium carbonate + titania
(3) Heat at 1000 degrees C
(4) ... 

**Synthesis Project API**
Query: “BaTiO₃”

**Embedding API**  **AVAILABLE**
“BaTiO₃” -> most_similar()
==> [“SrTiO₃”, “PbTiO₃”, ...]

**Aggregate API**  **AVAILABLE**
“Barium titanate”
Top co-occurring mat’ls: TiO₂, ...
Common topics: piezoelectric, ...
Example: suggesting synthesis conditions for a specific morphology in titania

Experimentally-accessible (and reported) variables to facilitate practical synthesis route planning.

Edward Kim et al., Chemistry of Materials 2017
Virtual synthesis screening is hard: data is sparse & scarce

“Sparse” = high-dimensional vector of synthesis actions
“Scarce” = materials of interest → not many papers published to train on

Can deep learning / generative models be useful for synthesis screening?
Variational autoencoder:
- Loss = reconstruction + f(Gaussian)
- Also a generative model

Edward Kim et al., npj Computational Materials 2017

Collaborator, Stefanie Jegelka, CSAIL, MIT
Data augmentation with text & data mining

Edward Kim et al., *npj Computational Materials* 2017
Example: suggesting synthesis conditions for stabilizing desired materials

Polymorphs for MnO$_2$ overlaid with most probable alkali-ion use in synthesis (intercalation-based phase stability)

Edward Kim et al., npj Computational Materials 2017
Exploratory: Rare phase in common material

Clustering of latent space shows driving conditions for polymorph of TiO$_2$ for photocatalysis

Edward Kim et al., *npj Computational Materials* 2017
Comparison of literature / virtual samples for SrTiO$_3$ synthesis

<table>
<thead>
<tr>
<th>Calcination</th>
<th>Sintering</th>
<th>Annealing</th>
<th>NaOH (M)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>800°C, 2h</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>Ye et al, 2016</td>
</tr>
<tr>
<td>800°C, 2h</td>
<td>1250°C, 2h</td>
<td>-</td>
<td>-</td>
<td>Zhao et al, 2004</td>
</tr>
<tr>
<td>1000°C, 12h</td>
<td>-</td>
<td>500°C, 2h</td>
<td>-</td>
<td>Zhao et al, 2015</td>
</tr>
<tr>
<td>600-750°C, 4h</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Puangpetch et al, 2008</td>
</tr>
<tr>
<td>721°C, 1.8h</td>
<td>-</td>
<td>468°C, 0.4h</td>
<td>-</td>
<td>N/A</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>450°C, 0.9h</td>
<td>1</td>
<td>N/A</td>
</tr>
<tr>
<td>955°C, 6h</td>
<td>1182°C, 7.5h</td>
<td>-</td>
<td>-</td>
<td>N/A</td>
</tr>
</tbody>
</table>

One cannot train a model exclusively on literature data and classify something as successful or not, since there are no negative examples in the literature

Edward Kim et al., *npj Computational Materials* 2017
Key elements in machine learning for energy technology

Ramprasad et al npj computational materials, 2017
Applicability and Next steps

• Continue to improve pipeline and disseminate information to the community
• Inform structure for data going forward
• Use cases in:
  • Solid state synthesis, hydrothermal and sol gel methods
  • Alloy design
  • Electrolyte performance
Thank you
olivetti.mit.edu
synthesisproject.org