MRS Open Data Challenge

2018 MRS Fall Meeting
November 25 – 30, 2018
OPEN DATA CHALLENGE

Contribute to the materials science and informatics communities by completing one or more of the following:

1. Curating or developing your own materials dataset.
2. Applying materials informatics techniques to your dataset to learn something new.

Finally, build a presentation to share your methodology and results with the scientific community!
OUTLINE

1. Why participate in the challenge?
2. What can open data + ML enable in materials science?
3. Data-driven success stories
4. How can I contribute?
5. Challenge details
6. Q&A
WHY PARTICIPATE?
WHY PARTICIPATE?

• Promote and participate in good scientific citizenship.
• Build in-demand skills for your career.
• Innovate how materials research is carried out by bringing informatics techniques to practice.
69% increase in citations for papers with publicly available data vs. papers that did not publish data.

Source: PLOS one [https://doi.org/10.1371/journal.pone.0000308](https://doi.org/10.1371/journal.pone.0000308)
SKILLS

Statistical Analysis and Data Mining
2nd most in-demand skill in 2018

Source: LinkedIn Learning
“Access to high-quality data emerged as the most important need and one of the most difficult for [the materials and chemicals] industry to address”

Source: NIST – Economic Analysis of National Needs for Tech Infrastructure to Support the MGI
ML for Materials Science
WHY DO WE NEED ML & INFORMATICS FOR MATERIALS SCIENCE?

- Leverage data as it is collected
- Glue together heterogeneous sources of data
- Systematically explore high dimensional design spaces
- Accelerate materials discovery
RECIPE FOR SUCCESS: DOMAIN KNOWLEDGE + ML

• Why is domain knowledge important?
  • We don’t want nonsense predictions: physics must be obeyed
  • We don’t want to have to reinvent the wheel: use data to learn the hard parts
  • Circumvent costly experiments or computation

• Domain knowledge can come in many forms
  • Known correlations between properties (e.g. color and band gap)
  • Known analytical relations (e.g. Arrhenius relation)
  • Building design candidates to choose from
  • There is no single method of building in domain knowledge—creativity is required
Sequential (active) learning finds high performance candidates faster than benchmark.

Important: The goal is not to replace experiments altogether, but to reduce the number of experiments.
KEY TAKEAWAYS

• Machine learning can be used to accelerate materials development
• Materials science requires domain-specific machine learning approaches
• Collaboration between machine learning experts and materials scientists is critical

Traditional materials development

Today’s state-of-the-art material

With machine learning: fewer experiments

Next-generation target
DATA-DRIVEN SUCCESS STORIES
HIGH-T FERROELECTRIC PEROVSKITES

• **Design goal**: Maximize Curie Temp of perovskite systems, with structure constraint

• **Training dataset**: 167 structures, 117 $T_C$ values (exp. results)

• **ML approach**: SVC for structures; SVR with bootstrap UQ

---

PEROVSKITE CASE STUDY: WORKFLOW

PEROVSKITE CASE STUDY: RESULTS

• Search space contains 61,506 candidates
• 5 total iterations (2 materials each) after initial training set
• Identified 6 new perovskites
  • 2 rank in the top 5 for Tc incl. training data

• **Design goal:** Identify novel Heusler-structured compounds
• **Training dataset:** 1,948 ternary structures from Pearson and ASM
• **ML approach:** Random forest, site-specific featurization

HEUSLER CASE STUDY: WORKFLOW


Significant data extraction and cleaning effort

Crystallographic and chemical features

ML model > heuristic

Table 1. Descriptors Used in Machine-Learning Model for Heusler Compounds $AB_2C$

<table>
<thead>
<tr>
<th>descriptor</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. group number of element $B$</td>
<td>0.1718</td>
</tr>
<tr>
<td>2. total number of $p$ valence electrons</td>
<td>0.1568</td>
</tr>
<tr>
<td>3. radius difference $A/B$</td>
<td>0.0959</td>
</tr>
<tr>
<td>4. electronegativity value of element $B$</td>
<td>0.0543</td>
</tr>
<tr>
<td>5. group number of element $A$</td>
<td>0.0956</td>
</tr>
<tr>
<td>6. group number of element $C$</td>
<td>0.0466</td>
</tr>
<tr>
<td>7. radius difference $A/C$</td>
<td>0.0465</td>
</tr>
<tr>
<td>8. radius value of element $C$</td>
<td>0.0361</td>
</tr>
<tr>
<td>9. molar mass value of element $C$</td>
<td>0.0353</td>
</tr>
<tr>
<td>10. radius difference $B/C$</td>
<td>0.0343</td>
</tr>
<tr>
<td>11. total number of $d$ valence electrons</td>
<td>0.0317</td>
</tr>
<tr>
<td>12. molar mass value of element $B$</td>
<td>0.0284</td>
</tr>
<tr>
<td>13. electronegativity value of element $A$</td>
<td>0.0284</td>
</tr>
<tr>
<td>14. electronegativity value of element $C$</td>
<td>0.0281</td>
</tr>
<tr>
<td>15. radius value of element $B$</td>
<td>0.0266</td>
</tr>
<tr>
<td>16. total number of $s$ valence electrons</td>
<td>0.0251</td>
</tr>
<tr>
<td>17. total number of valence electrons</td>
<td>0.0243</td>
</tr>
<tr>
<td>18. molar mass value of element $C$</td>
<td>0.0195</td>
</tr>
<tr>
<td>19. period number of element $A$</td>
<td>0.0179</td>
</tr>
<tr>
<td>20. radius value of element $C$</td>
<td>0.0175</td>
</tr>
<tr>
<td>21. period number of element $B$</td>
<td>0.0119</td>
</tr>
<tr>
<td>22. period number of element $C$</td>
<td>0.0092</td>
</tr>
</tbody>
</table>
HEUSLER CASE STUDY: RESULTS

400k compositions screened
12/14 true positives
7/7 true negatives

Entire community finds
50-60 new Heuslers/year

DETAiLS

Create
Create or curate an open dataset

Label
Label the data set and data points with sufficient metadata and citations so someone else can understand how the data were generated.

Analyze
Analyze your data using materials informatics.

Describe
How did you analyze your data? What tools did you use? What did you learn?

Present
Create a poster, slide deck, Prezi, applet, or iPython notebook describing the steps necessary to implement your analysis and showcasing your results.

TOOLS

Data

Open Citrination, NIST Materials Data Repository, Google Dataset, DRYAD, Materials Data Facility, Research Articles, Nature Data, Materials Project

Machine Learning

UI Based

Citrination.com

Python Based

Scikit-learn, TensorFlow, Keras, Citrine PyCC API
WHAT IS MY ROLE?

Students looking to learn data engineering skills can serve as data managers, using **basic programming** to scrape and transform data.

Data Manager
- Aggregates and Transforms Data

Data Scientist
- Proposes, Designs, & Runs Analytics

Domain Expert
- Chooses Data Sources
- Explains Desired Optimization Target

Students can leverage their backgrounds by forming teams.

Chemist, metallurgist, material science, and other experimentalist can focus on using their domain expertise.

Physics, mathematics, computer science, and other computational students run analytics that to test hypothesis.
TIMELINE

December 13th
Kick-Off Webinar

December 14th
Submissions Open

January 15th
Materials Informatics Tools Webinar

February 15th
Deadline to Submit

March 15th
Finalist Selected and Notified

April 24th
Winners Announced at 2019 MRS Spring Meeting

1st Place 2nd Place 3rd Place
$1000 $500 $250
JUDGING

• Participants will be judged on:
  • the clarity of their communication
  • novelty of their dataset and/or analysis approach
  • explanation of their methodology
• Finalists selected by March 15, 2019.
• Judged by experts from the materials science and informatics communities.
• Rubrics will be shared via the MRS website and to students who sign up through the web portal.
THANK YOU!
THANK YOU!

Open Data Challenge: mrs.org/open-data-challenge
Data-driven materials newsletter: citrine.io/newsletter
Emails: Malcolm – mdavidson@citrine.io
        Josh – jtappan@citrine.io
## FOR REFERENCE: WHAT MAKES MACHINE LEARNING FOR MATERIALS UNIQUE?

<table>
<thead>
<tr>
<th></th>
<th>TRADITIONAL AI APPLICATIONS</th>
<th>MATERIALS AI APPLICATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA TYPE</td>
<td>Sometimes heterogeneous</td>
<td>Highly heterogeneous, requires extensive custom feature engineering and domain knowledge</td>
</tr>
<tr>
<td>DATA VOLUME</td>
<td>Big, dense</td>
<td>Small, sparse</td>
</tr>
<tr>
<td>ESTABLISHED DOMAIN KNOWLEDGE</td>
<td>Not applicable—rely on data to learn patterns</td>
<td>Must be physics-aware</td>
</tr>
<tr>
<td>UNCERTAINTY QUANTIFICATION</td>
<td>Usually unimportant</td>
<td>Always important</td>
</tr>
<tr>
<td>INTERPRETABILITY</td>
<td>Usually unimportant</td>
<td>Often required by scientists &amp; engineers</td>
</tr>
<tr>
<td>PREDICTION TASK</td>
<td>Pattern-match common cases</td>
<td>Discover high-performance materials</td>
</tr>
</tbody>
</table>