MRS Open Data Challenge

2018 MRS Fall Meeting

November 25 – 30, 2018

CITRINE INFORMATICS

Josh Tappan

Community

Malcolm Davidson

Open Data

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

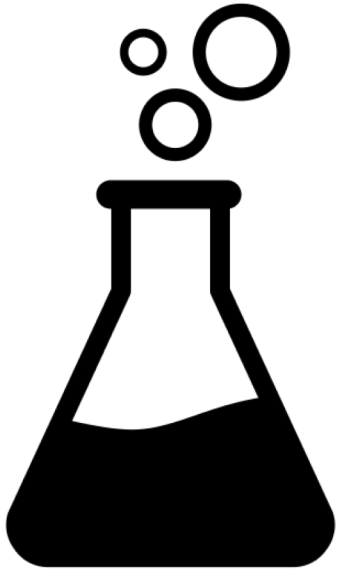
1. Why participate in the challenge?
2. What can open data + ML enable in materials science?
3. Data-driven success stories
4. How can I contribute?
5. Challenge details
6. Q&A

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.

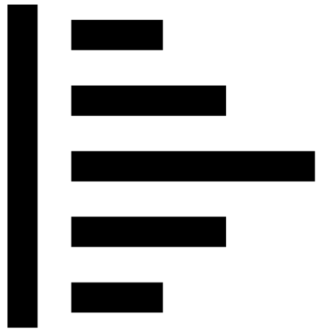
IMPACT



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

SKILLS

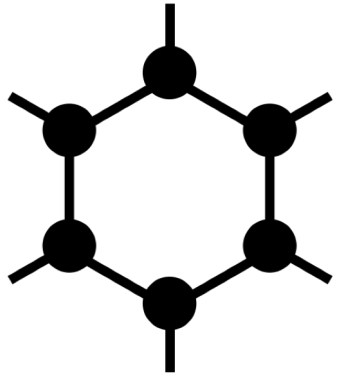


Statistical Analysis and Data Mining

2nd most in-demand skill in 2018

Source: LinkedIn Learning

MATERIALS INNOVATION



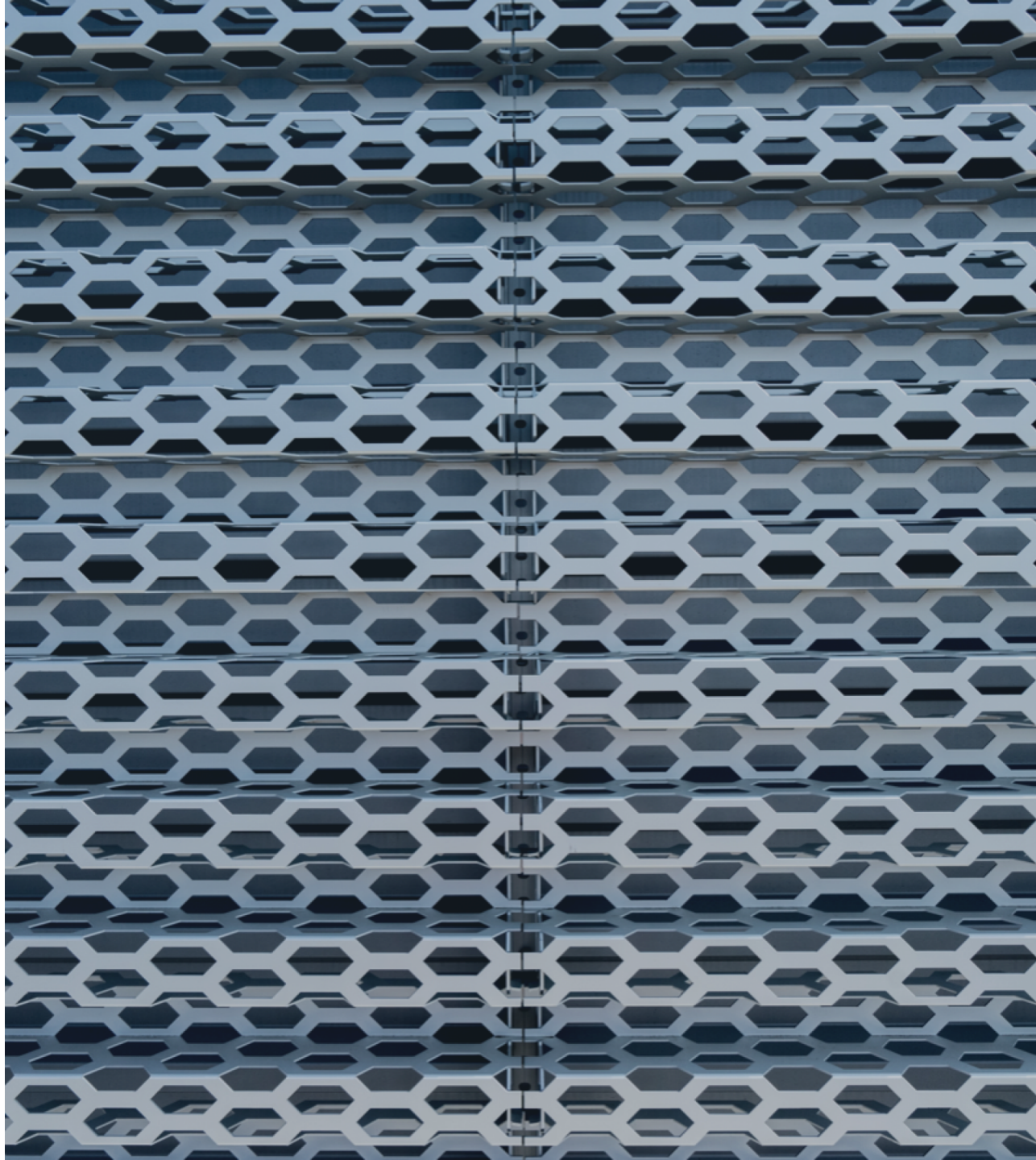
"Access to high-quality data emerged as the most important need and one of the most difficult for [the materials and chemicals] industry to address"

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

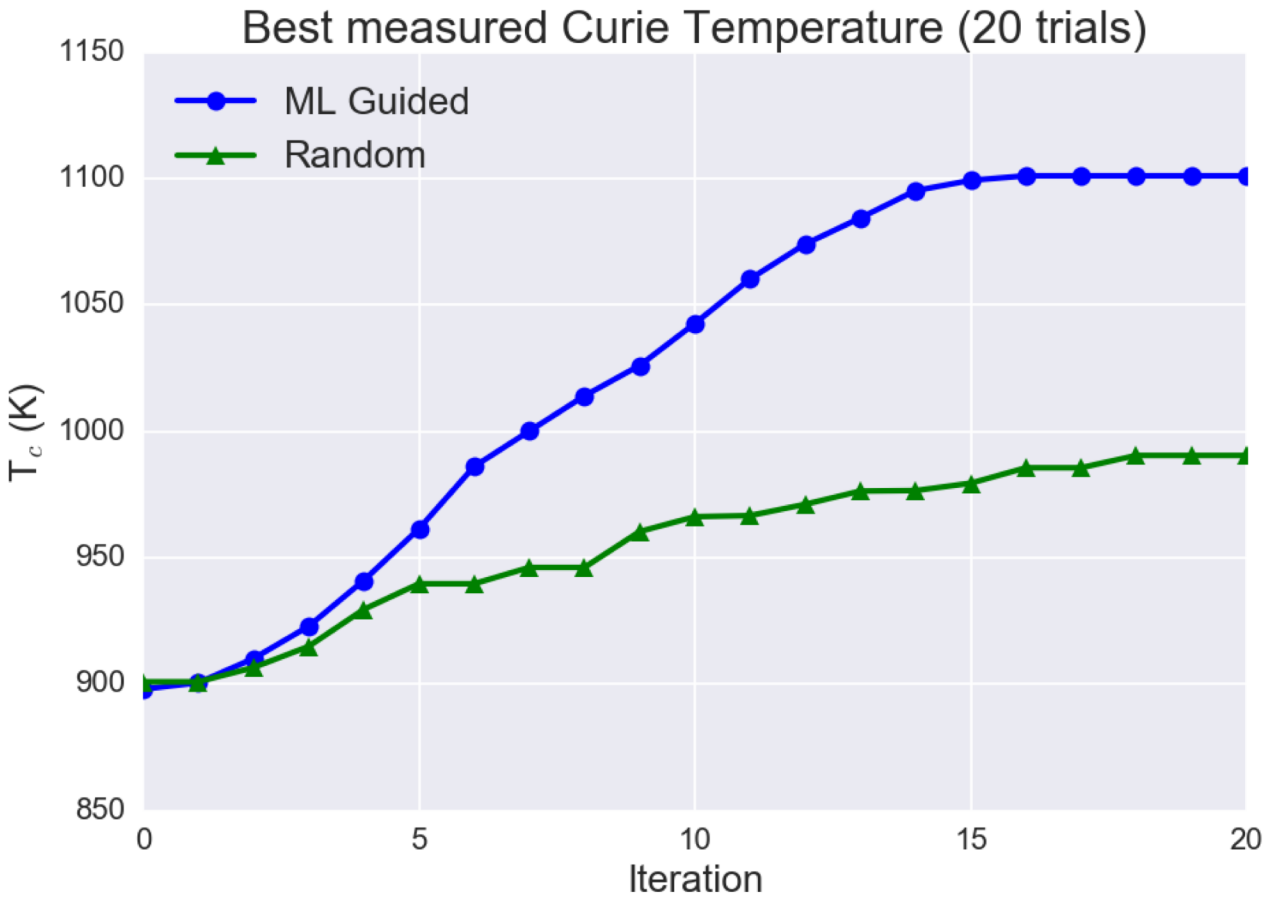


RECIPE FOR SUCCESS: DOMAIN KNOWLEDGE + ML

- 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 LEARNING

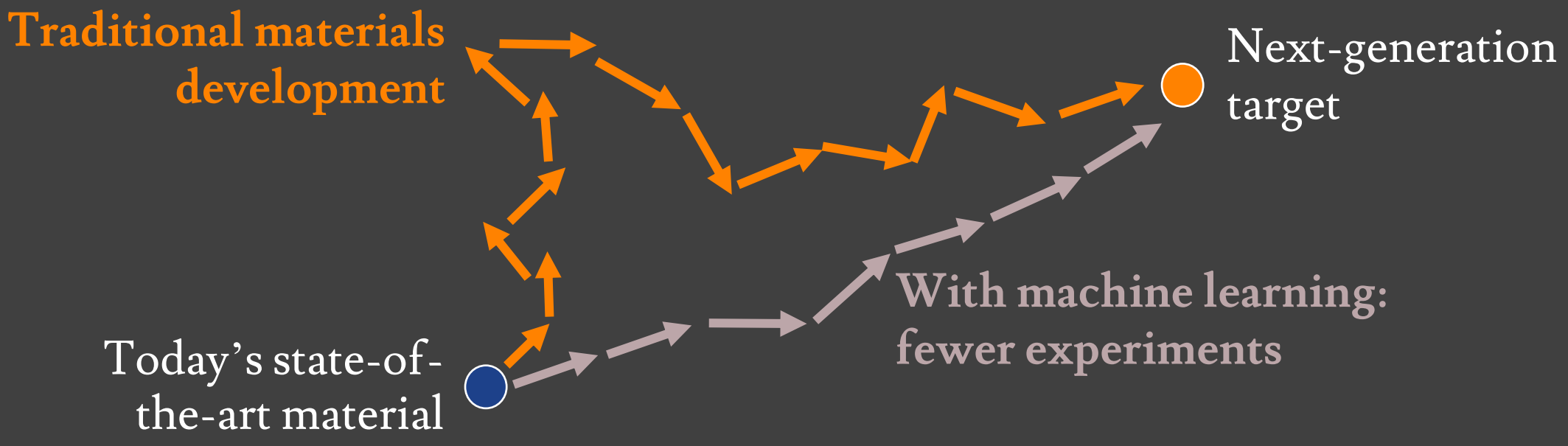


Sequential (active) learning finds high performance candidates faster than benchmark

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

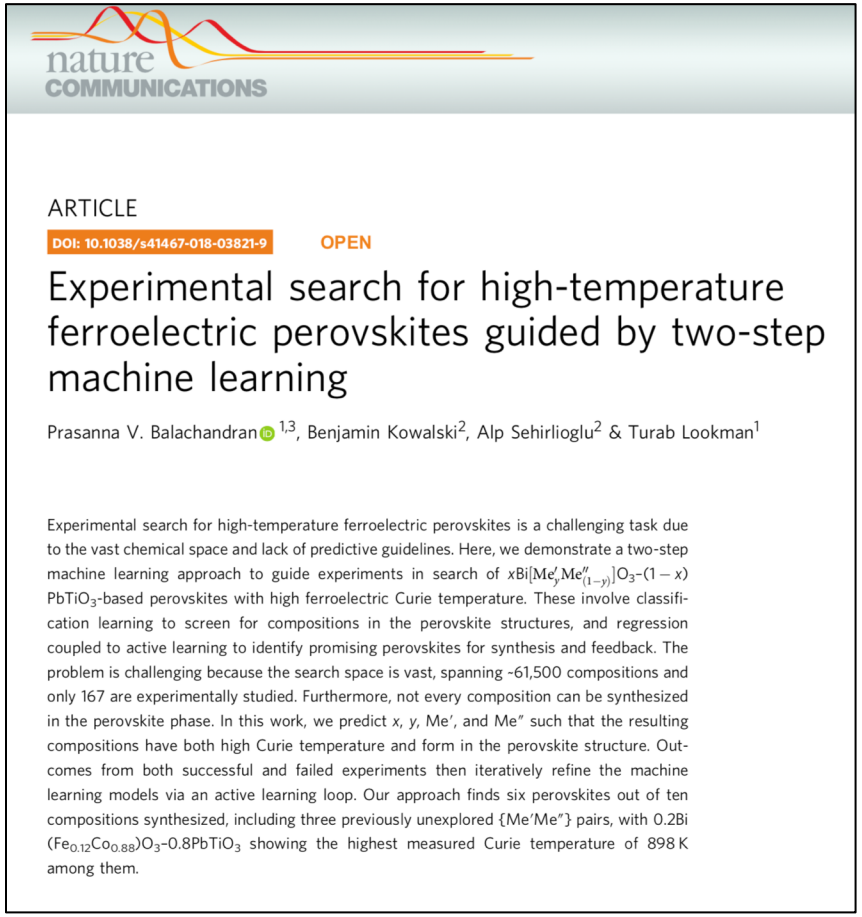
KEY TAKEAWAYS

- 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



DATA-DRIVEN SUCCESS STORIES

HIGH-*T* FERROELECTRIC PEROVSKITES



nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-03821-9 OPEN

Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning

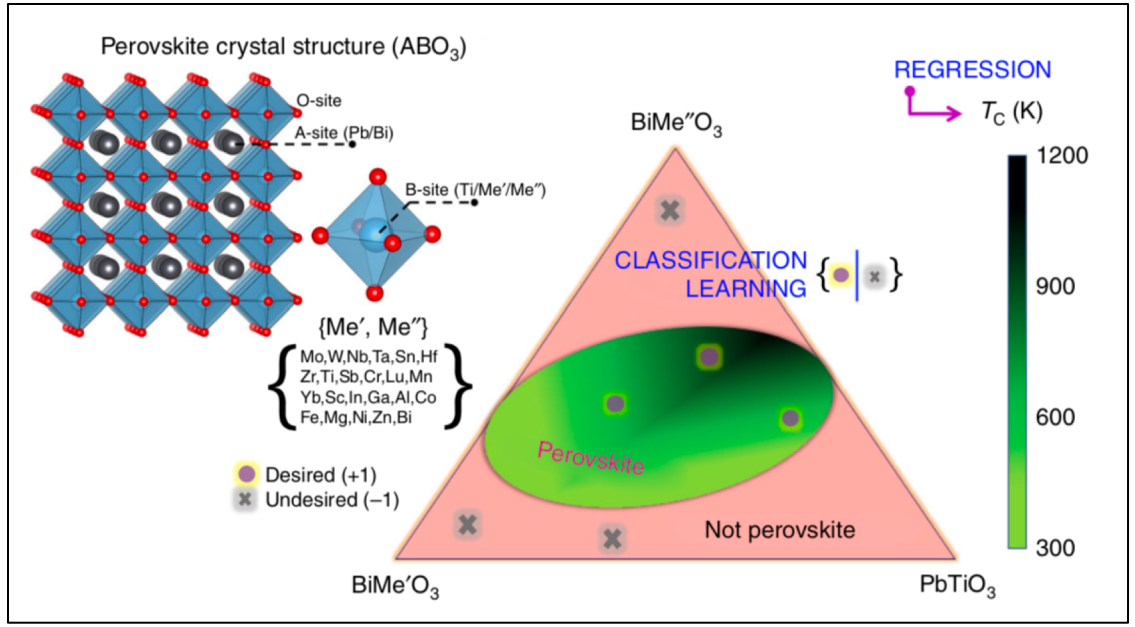
Prasanna V. Balachandran[1,3], Benjamin Kowalski[2], Alp Sehirlioglu[2] & Turab Lookman[1]

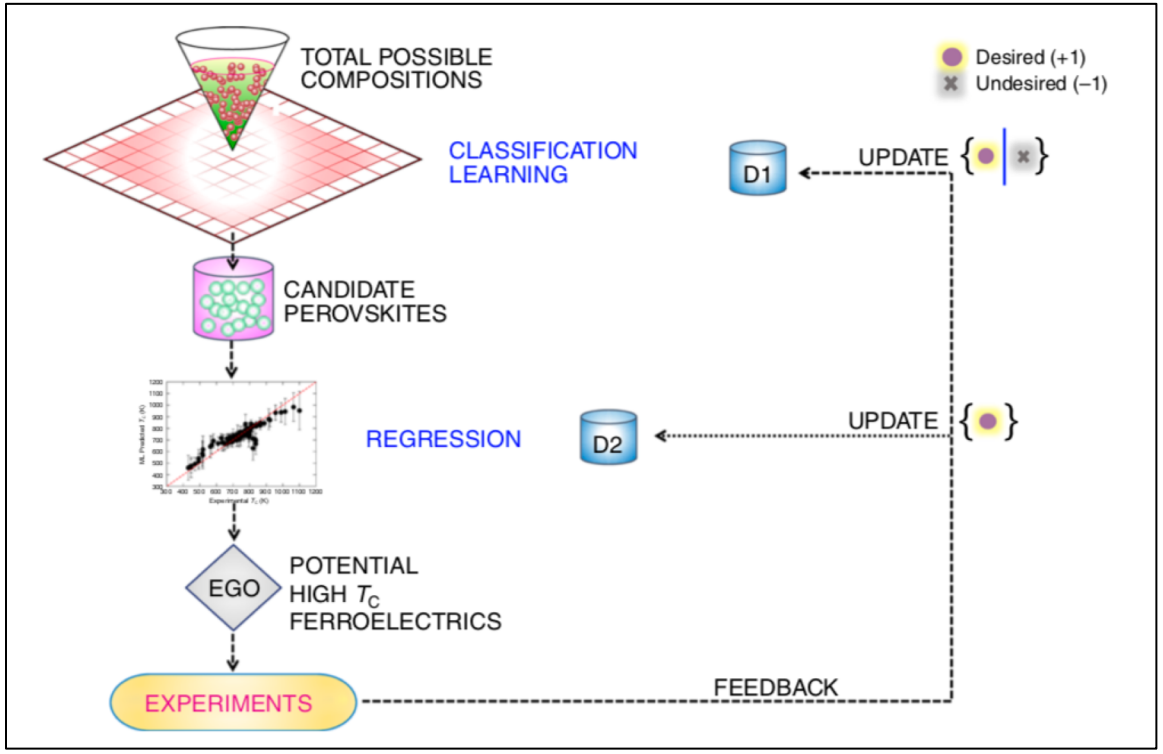
Experimental search for high-temperature ferroelectric perovskites is a challenging task due to the vast chemical space and lack of predictive guidelines. Here, we demonstrate a two-step machine learning approach to guide experiments in search of $xBi[Me'_yMe''_{(1-y)}]O_3$-$(1-x)PbTiO_3$-based perovskites with high ferroelectric Curie temperature. These involve classification learning to screen for compositions in the perovskite structures, and regression coupled to active learning to identify promising perovskites for synthesis and feedback. The problem is challenging because the search space is vast, spanning ~61,500 compositions and only 167 are experimentally studied. Furthermore, not every composition can be synthesized in the perovskite phase. In this work, we predict *x*, *y*, Me′, and Me″ such that the resulting compositions have both high Curie temperature and form in the perovskite structure. Outcomes from both successful and failed experiments then iteratively refine the machine learning models via an active learning loop. Our approach finds six perovskites out of ten compositions synthesized, including three previously unexplored {Me′Me″} pairs, with $0.2Bi(Fe_{0.12}Co_{0.88})O_3$-$0.8PbTiO_3$ showing the highest measured Curie temperature of 898 K among them.

- **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

Reference: Balachandran, P. V., Kowalski, B., Sehirlioglu, A., & Lookman, T. (2018). Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. *Nature Communications*, 9.

PEROVSKITE CASE STUDY: WORKFLOW





Reference: Balachandran, P. V., Kowalski, B., Sehirlioglu, A., & Lookman, T. (2018). Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. *Nature Communications*, 9.

PEROVSKITE CASE STUDY: RESULTS

Table 1 List of ten new compositions predicted and validated by ML and experimental measurements

Iteration #	Compositions	Perovskites (+1 or −1)		Curie temperature T_C (K)			
		Prediction from SVC_{rbf}	XRD measurements	Prediction from SVR_{rbf} ($\widehat{T_C} \pm \sigma$)	Dielectric measurements	EGO or Exploitation	*c/a*
1	$0.6Bi(Sc_{0.17}Ga_{0.83})O_3$-0.4PT	+1	−1	747 ± 223	-	EGO	-
2a	$0.35Bi(In_{0.31}Sc_{0.69})O_3$-0.65PT	+1	+1	923 ± 343	733	EGO	1.029
2b	$0.5Bi(Sc_{0.56}Ga_{0.44})O_3$-0.5PT	+1	−1	846 ± 229	-	EGO	-
3a	$0.25Bi(Sc_{0.47}Ga_{0.53})O_3$-0.75PT	+1	+1	778 ± 135	798	EGO	1.058
3b	$0.7Bi(Fe_{0.73}Co_{0.27})O_3$-0.3PT	+1	−1	929 ± 94	-	Exploitation	-
4a	$0.4Bi(Fe_{0.19}Sc_{0.81})O_3$-0.6PT	+1	+1	749 ± 161	728	EGO	1.026
4b	$0.2Bi(Fe_{0.12}Co_{0.88})O_3$-0.8PT	+1	+1	815 ± 24	898	Exploitation	1.066
5a	$0.3Bi(Yb_{0.44}Al_{0.56})O_3$-0.7PT	+1	−1	801 ± 125	-	EGO	-
5b	$0.3Bi(Ni_{0.50}Sn_{0.50})O_3$-0.7PT	+1	+1	780 ± 123	658	EGO	1.027
5c	$0.2Bi(Co_{0.90}Al_{0.10})O_3$-0.8PT	+1	+1	843 ± 33	883	Exploitation	1.063

EGO or Exploitation within the active learning strategy recommended the corresponding composition for synthesis. SVC_{rbf} and SVR_{rbf} are the classification learning and regression ML methods, respectively. Labels +1 and −1 in the XRD measurements column refer to compositions in the desired perovskite phase and those that contained undesired secondary phases, respectively. The XRD and dielectric measurements provide the ground truth for validating the SVC_{rbf} and SVR_{rbf} predictions. The tetragonality (*c/a* ratio) from XRD measurements are also given. In the column named "Iteration", we use a, b, or c as indicators to only uniquely identify compositions and they carry no scientific information

- 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

Reference: Balachandran, P. V., Kowalski, B., Sehirlioglu, A., & Lookman, T. (2018). Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. *Nature Communications*, 9.

SCREENING FOR HEUSLER ALLOYS

CHEMISTRY OF MATERIALS

Article

pubs.acs.org/cm

High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds

Anton O. Oliynyk,[†] Erin Antono,[‡] Taylor D. Sparks,[§] Leila Ghadbeigi,[§] Michael W. Gaultois,[¶] Bryce Meredig,[*,‡] and Arthur Mar[*,†]

[†]Department of Chemistry, University of Alberta, Edmonton, AB T6G 2G2 Canada

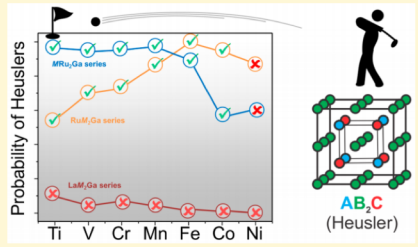
[‡]Citrine Informatics, Redwood City, California 94063 United States

[§]Department of Materials Science and Engineering, University of Utah, Salt Lake City, Utah 84112 United States

[¶]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW United Kingdom

Supporting Information

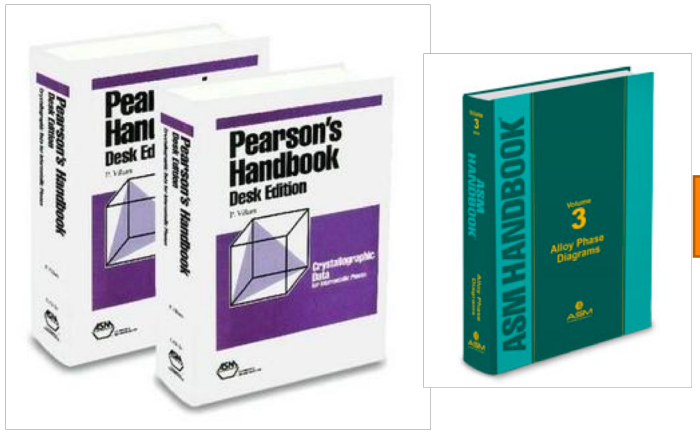
ABSTRACT: A machine-learning model has been trained to discover Heusler compounds, which are intermetallics exhibiting diverse physical properties attractive for applications in thermoelectric and spintronic materials. Improving these properties requires knowledge of crystal structures, which occur in three subtle variations (Heusler, inverse Heusler, and CsCl-type structures) that are difficult, and at times impossible, to distinguish by diffraction techniques. Compared to alternative approaches, this Heusler discovery engine performs exceptionally well, making fast and reliable predictions of the occurrence of Heusler vs non-Heusler compounds for an arbitrary combination of elements with no structural input on over 400 000 candidates. The model has a true positive rate of 0.94 (and false positive rate of 0.01). It is also valuable for data sanitizing, by flagging questionable entries in crystallographic databases. It was applied to screen candidates with the formula AB_2C and predict the existence of 12 novel gallides MRu_2Ga and RuM_2Ga (M = Ti−Co) as Heusler compounds, which were confirmed experimentally. One member, $TiRu_2Ga$, exhibited diagnostic superstructure peaks that confirm the adoption of an ordered Heusler as opposed to a disordered CsCl-type structure.



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

Reference: Oliynyk, A. O., Antono, E., Sparks, T. D., Ghadbeigi, L., Gaultois, M. W., Meredig, B., & Mar, A. (2016). High-throughput machine-learning-driven synthesis of full-Heusler compounds. *Chemistry of Materials*, 28(20), 7324-7331.

HEUSLER CASE STUDY: WORKFLOW

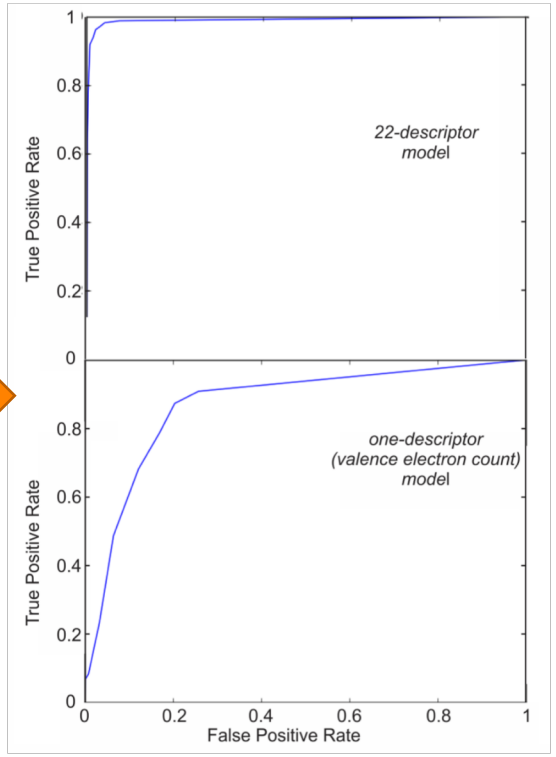


Significant data extraction and cleaning effort

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

descriptor	weight
1. group number of element *B*	0.1718
2. total number of p valence electrons	0.1568
3. radius difference *A*/*B*	0.0959
4. electronegativity value of element *B*	0.0543
5. group number of element *A*	0.0536
6. group number of element *C*	0.0466
7. radius difference *A*/*C*	0.0465
8. radius value of element *C*	0.0361
9. molar mass value of element *C*	0.0353
10. radius difference *B*/*C*	0.0343
11. total number of d valence electrons	0.0317
12. molar mass value of element *B*	0.0284
13. electronegativity value of element *A*	0.0284
14. electronegativity value of element *C*	0.0281
15. radius value of element *B*	0.0266
16. total number of s valence electrons	0.0251
17. total number of valence electrons	0.0243
18. molar mass value of element *C*	0.0195
19. period number of element *A*	0.0179
20. radius value of element *C*	0.0175
21. period number of element *B*	0.0119
22. period number of element *C*	0.0092

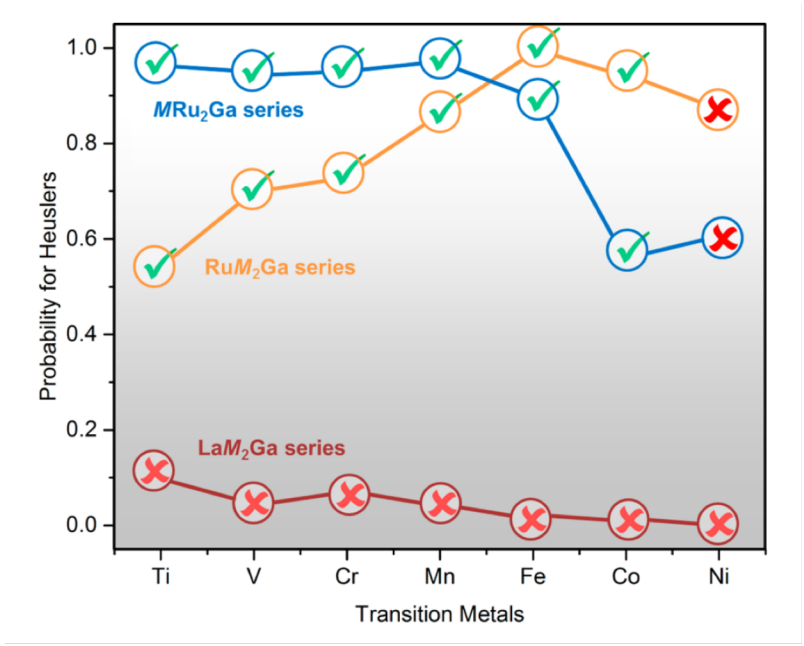
Crystallographic and chemical features



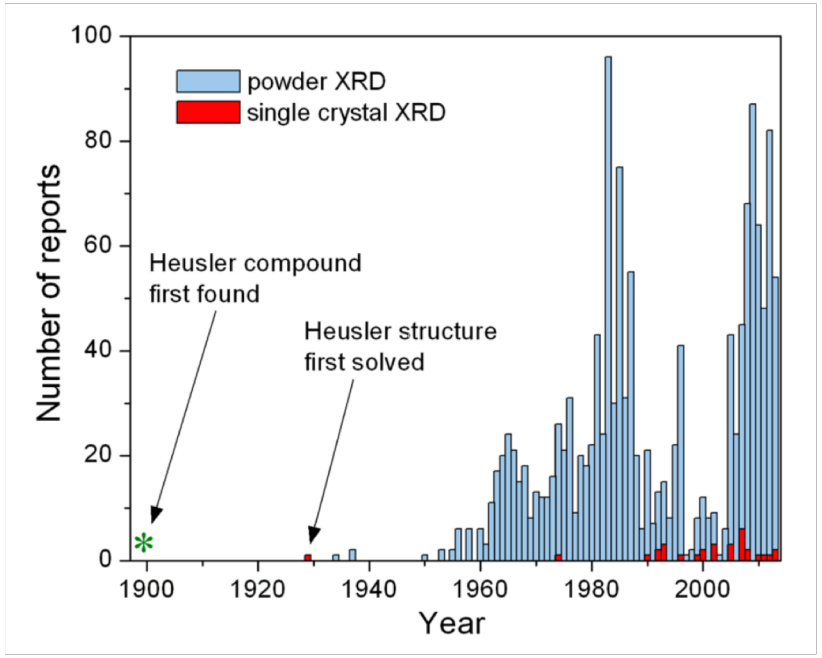
ML model > heuristic

Reference: Oliynyk, A. O., Antono, E., Sparks, T. D., Ghadbeigi, L., Gaultois, M. W., Meredig, B., & Mar, A. (2016). High-throughput machine-learning-driven synthesis of full-Heusler compounds. *Chemistry of Materials*, 28(20), 7324-7331.

HEUSLER CASE STUDY: RESULTS



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



Entire community finds
50-60 new Heuslers/year

Reference: Oliynyk, A. O., Antono, E., Sparks, T. D., Ghadbeigi, L., Gaultois, M. W., Meredig, B., & Mar, A. (2016). High-throughput machine-learning-driven synthesis of full-Heusler compounds. *Chemistry of Materials*, 28(20), 7324-7331.

CHALLENGE DETAILS

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

Students can leverage their backgrounds by forming teams

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

Data Scientist
- Proposes, Designs, & Runs Analytics

Domain Expert
- Chooses Data Sources
- Explains Desired Optimization Target

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!

CITRINE INFORMATICS

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

CITRINE INFORMATICS

FOR REFERENCE: WHAT MAKES MACHINE LEARNING FOR MATERIALS UNIQUE?

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