Entropy Targeted Active Learning for Bias Mitigation in Materials Data

Hengrui Zhang[1], Wei (Wayne) Chen[1], James M. Rondinelli[2], Wei Chen[1]

[1]Department of Mechanical Engineering, [2]Department of Materials Science & Engineering, Northwestern University

Overview

Data-centric informatics has promoted materials discovery and design

- Models, e.g., machine learning (ML), draw much attention
- Quality of data is equally important but less studied

This work focuses on data bias: lower bias → better coverage of design space → better generalizability of models

- A metric for measuring bias in materials data
- A framework guiding data acquisition to mitigate the bias

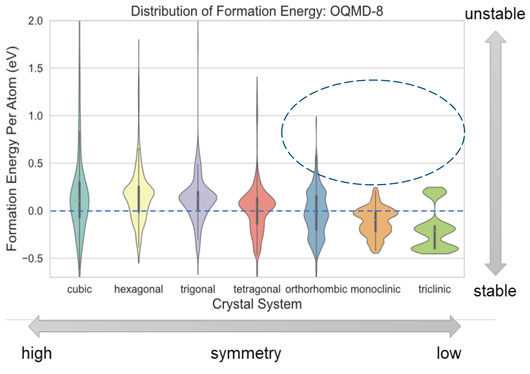
Problem Statement

Where materials informatics researchers get data

Literature	Databases	Experiments/computation
Usually "good performers" in certain applications	Built upon known structures/prototypes	Generate new data, can mitigate the bias

These data sources are often biased

Example: structure–stability bias in OQMD



$-\Delta E$ ~ stability

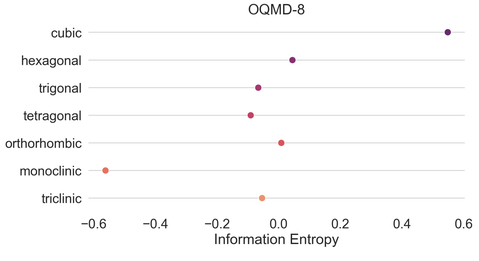
In physics:

symmetry↑ stability↑

While in the data:

The opposite.

Some crystal families are underrepresented.



Quantitative bias metric

Information entropy

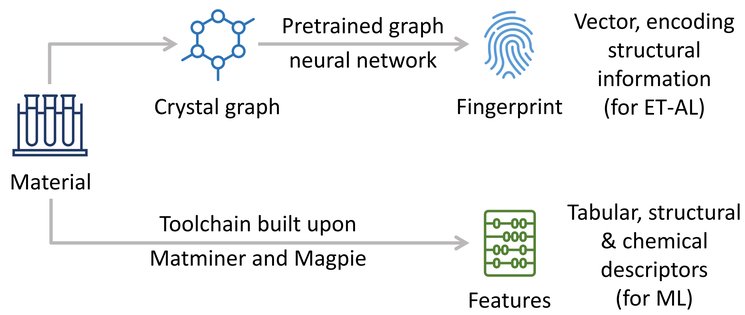
$$H(Y) = -\int_Y p(y) \ln p(y) \mathrm{d}y$$

$H(\Delta E)$ of a family ~ diversity

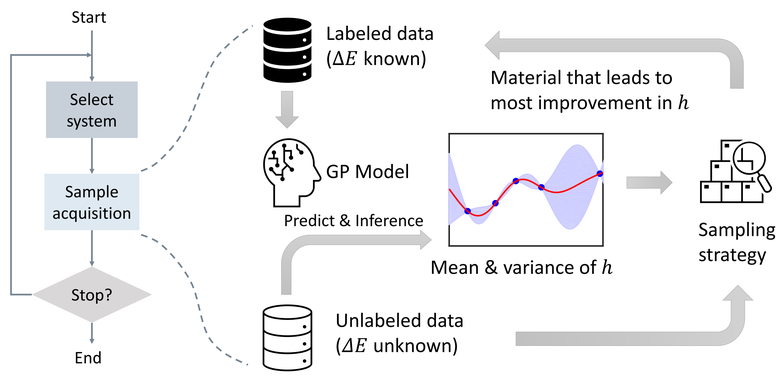
Fairness in diversity ~ bias

Methods

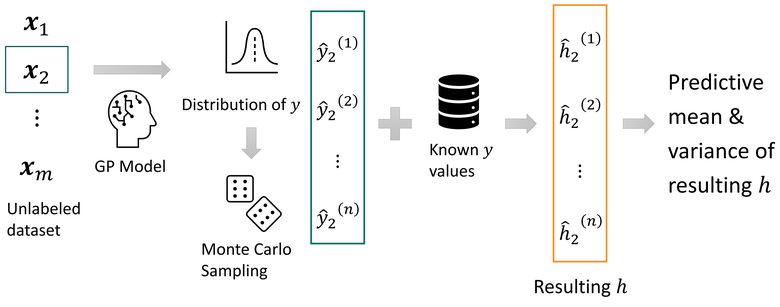
Representations of materials (input to the models)



ET-AL framework: maximize H of underrepresented systems



Monte Carlo inference for uncertainty in H



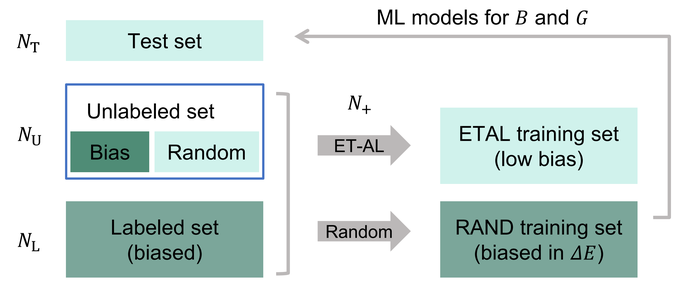
Outlook

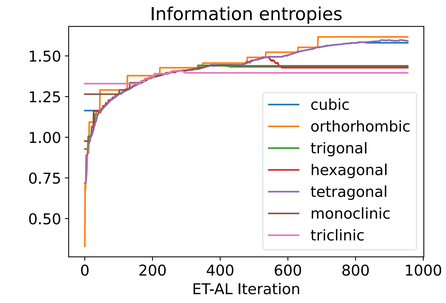
ET-AL can mitigate data bias, thus benefiting data-driven models

- Help researchers improve the quality of datasets
- Guide the construction of materials data platforms
- Applicable to data-centric informatics in other scientific domains

Results

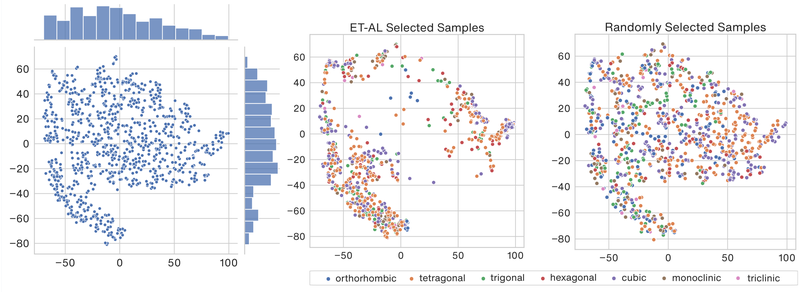
Virtual experiments on Jarvis dataset (size ~12K)



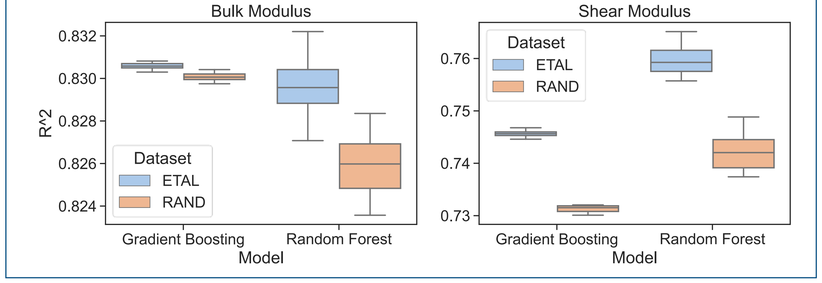


ET-AL successfully fixed the artificially created bias.

ET-AL selects samples in underrepresented regions



Reducing bias benefits **ML modeling** for materials properties.



Reference: H. Zhang, W. W. Chen, J. M. Rondinelli, & W. Chen. ET-AL: Entropy Targeted Active Learning for Bias Mitigation in Materials Data. *In preparation*.

This work was performed under the financial assistance award 70NANB19H005 from U.S. Department of Commerce, National Institute of Standards and Technology as a part of the Center for Hierarchical Materials Design (CHiMaD).