

# N

# Data Bias Detection and Mitigation in Materials Databases

Adetoye Adekoya, Whitney Tso, Hengrui Zhang, Wei Chen\*, David Seidman\*, Jeffrey Snyder\*

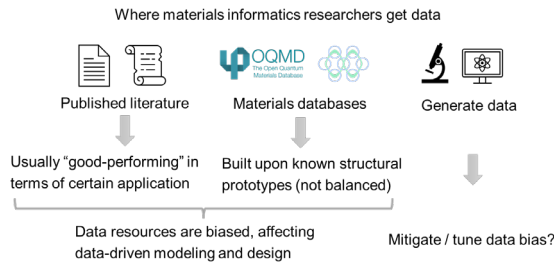
Academic Disciplines: Materials Science & Engineering, Mechanical Engineering

\*faculty advisors

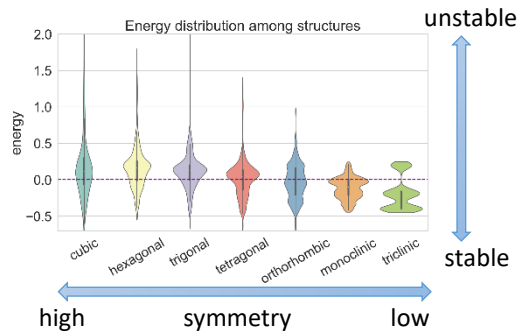
## Objective

Data bias is a ubiquitous but often overlooked problem that exists in most materials databases. Our objective is to develop computational methods that quantify the level of bias in datasets and mitigate bias by suggesting new data entries. This method aims to improve the quality of data to facilitate predictive modeling and design of materials.

## Motivation



Example: structure–stability bias among 8 common metal elements in OQMD



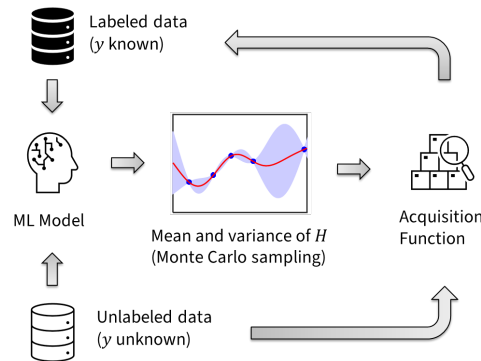
## Method

Information entropy

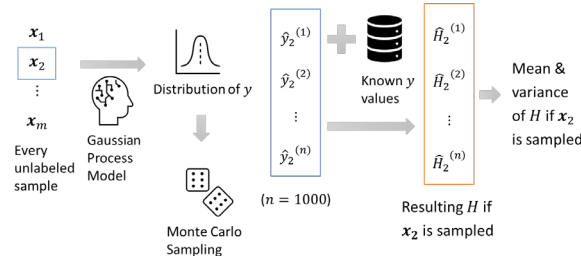
$$H(Y) = - \int_Y p(y) \ln p(y) dy$$

Quantifies the diversity of  $Y \rightarrow$  bias metric

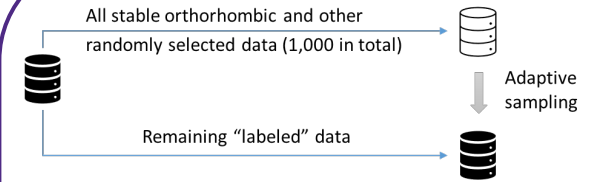
Maximizing  $H(\Delta E)$  in all crystal systems will reduce bias  $\rightarrow$  Adaptive Sampling



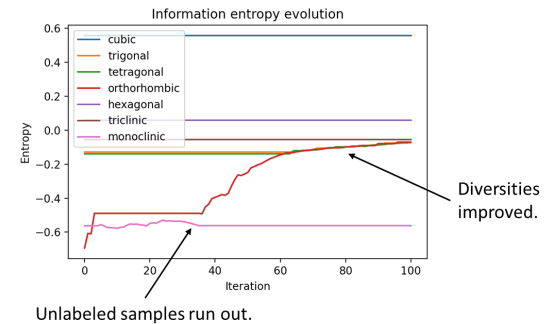
Monte Carlo sampling  $\rightarrow$  uncertainty in  $H$ :



## Results



$H(\Delta E)$  change in sampling process



Successfully fixed artificially created bias.

## Future Work

- Examine the effect of structure–stability bias on modeling other properties.
- Apply the method to guide materials database construction
- Tune data bias according to the need

