Machine Learned Interatomic Potentials with Active Learning

Graduate Student Fellows:
VISHU GUPTA
SEAN KOYAMA

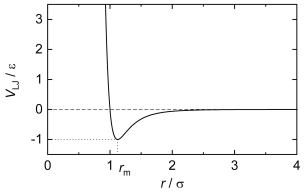
Faculty Advisors:
ANKIT AGRAWAL
JAMES RONDINELLI

Academic Disciplines:
ELECTRICAL & COMPUTER ENGINEERING
MATERIALS SCIENCE & ENGINEERING

June 10, 2021

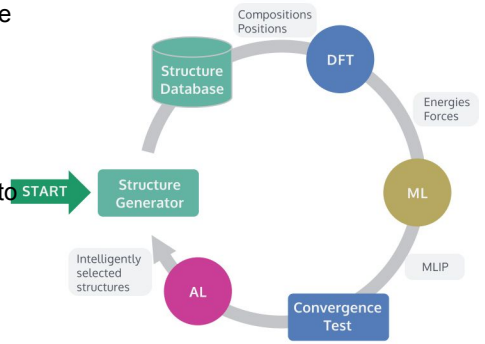
Research Outline

Machine learned interatomic potentials (MLIPs) can be fine-tuned to a specific system of interest and can be more accurate than empirical potentials. The training data is typically computed with density functional theory or a similar ab-initio method.



Problem: computing these MLIPs can be computationally intensive and inefficient, largely due to the inability to predict the minimal required training dataset for a desired level of accuracy.

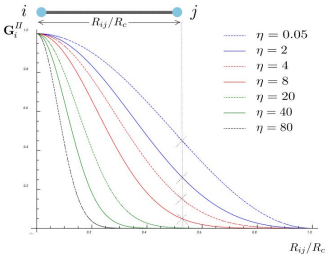
Solution: utilize an active learning loop to intelligently select input datasets in batches, reducing the total computational effort needed to train the MLIPs.



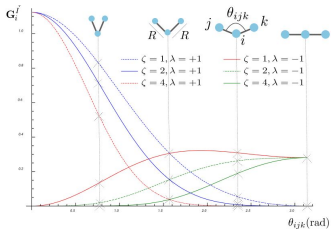
Structure and Descriptor Generation

To use structures as inputs into the machine learning model, we need to set descriptors for the structures - these are used to produce feature vectors which serve as surrogates to describe the structure.

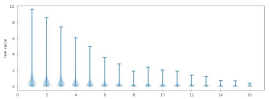
We use 2-body and 3-body (to capture the physics of silicon) atom-centered Gaussian descriptor set.



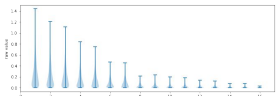
Evaluating descriptors: it can be done by testing different descriptor sets on the same test data, but this can be time consuming. Hence, for our initial analysis, we create violin plots to see how much each descriptor is being activated within our descriptor set.



Si structure set 1 (scaled equilibrium structures)



Si structure set 2 (random structures)

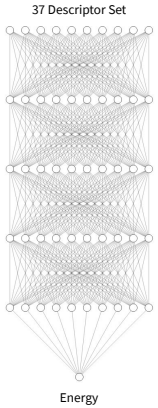


Active Learning Framework

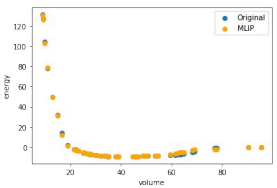
We perform active learning of regression models using expected model change (EMC) and query by committee (QBC) methods.

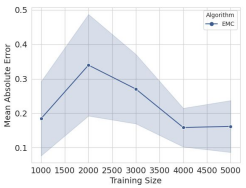
These algorithm assumes that you have a pool of unlabeled data points and a limited budget to label them. Thus, it combine the efficiency of the Machine Learning (ML) models with active learning approach to suggest optimal number of calculations to provide labeled data.

Reproducing E-V curve: Reproducing the energy-volume (E-V) curve is a minimum requirement for any interatomic potential. We successfully produced a model that was able to reproduce this curve.



AL Training Analysis: The model is able to predict the correct trend with only 200 volume series structure Here, active learning approach is used to achieve lower prediction errors by choosing less but more informative data points.

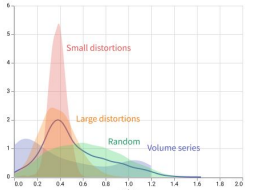




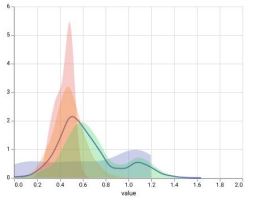
Descriptor Analysis and Future Goal

We can evaluate the active learning model by looking at the distribution of descriptor activation for each descriptor across the dataset for each of the structure type (Volume series, Small distortions, Large distortions, Random) individually.
Since active learning occurs in iterations, we can look at the descriptor distributions in the structure sets chosen in each active learning iteration.

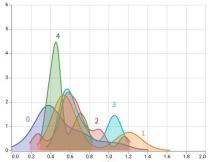
Full dataset



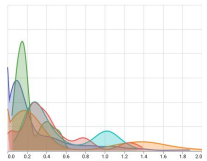
AL dataset



2-body descriptor



3-body descriptor



Future Goals: Structure selection and generation through active learning:

- Pure exploration (i.e. choosing points of maximum uncertainty)
- bayesian optimization with appropriate estimate of model improvement. etc.

NORTHWESTERN UNIVERSITY