

MATERIALS ENGINEERING

SEMINAR

“Semi-Supervised Machine Learning Methods and their Applications to Materials Informatics”

By

Kat Nykiel

Purdue MSE Preliminary Exam

Advisor: Professor Alejandro Strachan

ABSTRACT

Machine learning is a powerful tool with the potential to circumvent complex physical problems by mapping directly from inputs to outputs. In datasets where labels are abundant, supervised models are trained on the labeled data and used to infer labels for unlabeled samples. When only unlabeled samples are present, unsupervised methods attempt to identify interesting patterns in the data. However, when only a small portion of the data is labeled, a new set of tools emerges collectively known as semi-supervised learning. These methods use context from unlabeled data to better inform the model, typically with higher accuracy than pure supervised learning. The trend of large, unlabeled datasets with sparse labels is common to many problems in materials science, where acquiring new labels is often expensive. For example, semi-supervised learning is beneficial for predicting new materials, as this is a vast domain with a large amount of unlabeled data and few positive samples. However, the overlap of semi-supervised learning and materials science is limited. This paper seeks to review the assumptions and classes of semi-supervised learning methods and conduct a demonstration of how they might be applied to a materials problem. A self-training, pseudo-label approach is applied to a dataset of density-functional theory calculations to predict formation energy from a small amount of labeled data.

Date: Friday, December 9, 2022

Time: 8:30 am

Place: DLRC 143 B or WebEx via <https://purdue.webex.com/join/strachan>



School of Materials Engineering