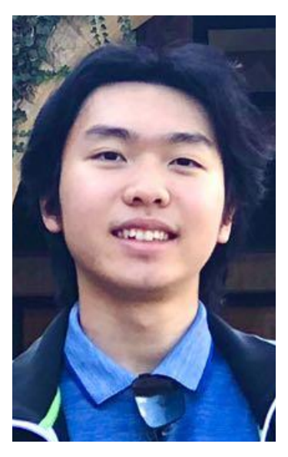


Student:
William Cai

Expanding the Spectrum of Synthesizable Materials Using Positive-Unlabeled Learning Model with Group-Based Sampling Method

William Cai[1], Evan Antoniuk[1], Evan Reed[1]

[1]Department of Materials Science and Engineering, Stanford University

Contact: willcai1@stanford.edu

Advisor:
Prof. Evan Reed

Mentor:
Evan Antoniuk

One of the largest outstanding problems in computational material science is to elucidate the full spectrum of materials that can be synthesized. The goal of the project is to develop a semi-supervised model for material synthesizability prediction. This project trains and evaluates a positive-unlabeled learning model with data from the Inorganic Crystal Structure Database (ICSD) and Materials Project Database.

Introduction

Background & Problems

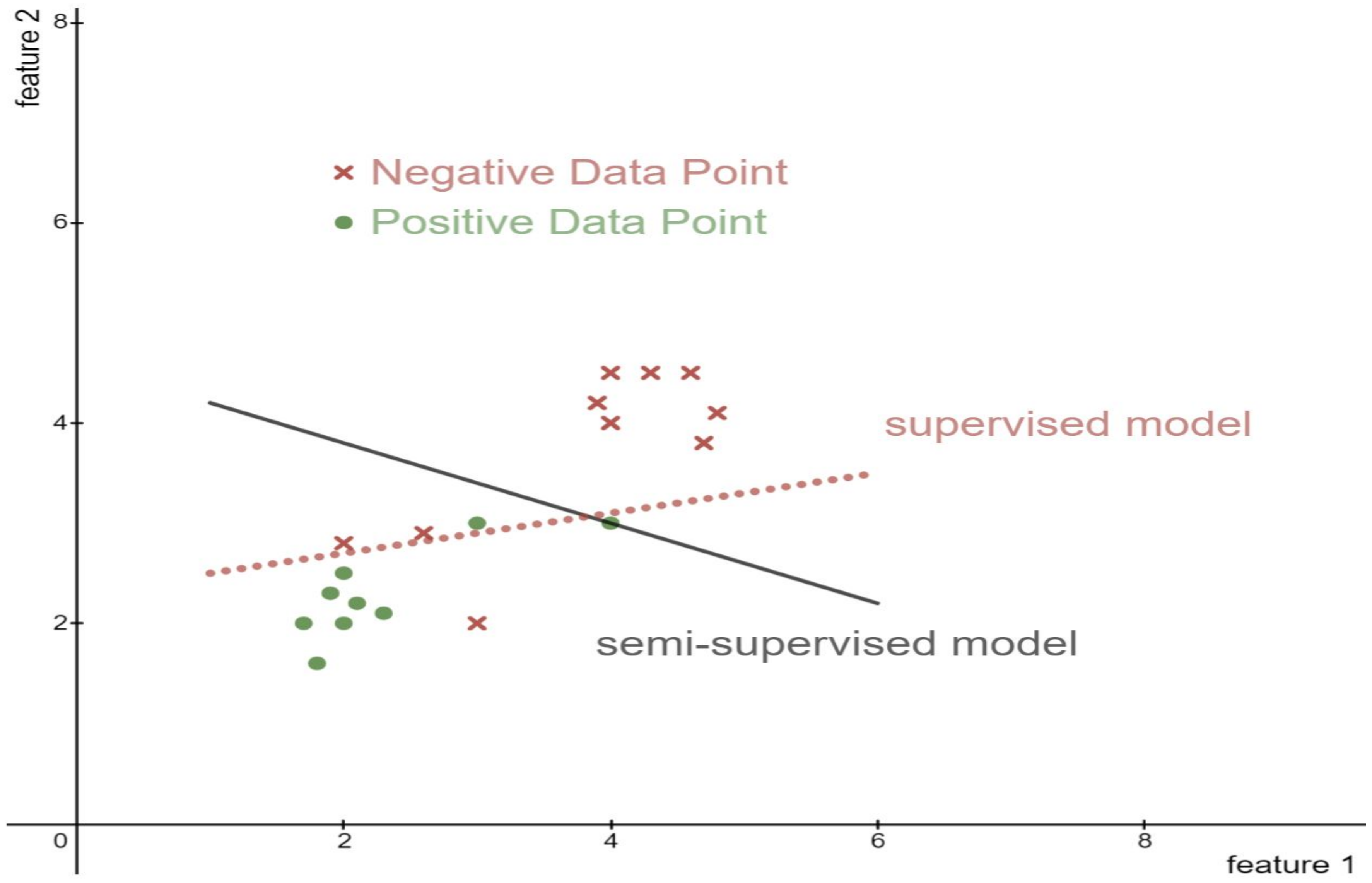
- Traditional approaches are usually very computationally expensive and limit the diversity of materials to be identified.
- Some material synthesis processes are costly and involve the use of hazardous chemicals or the production of hazardous byproducts (Szczęśniak et al., 2020).

Current Stage of Data-Driven Approaches

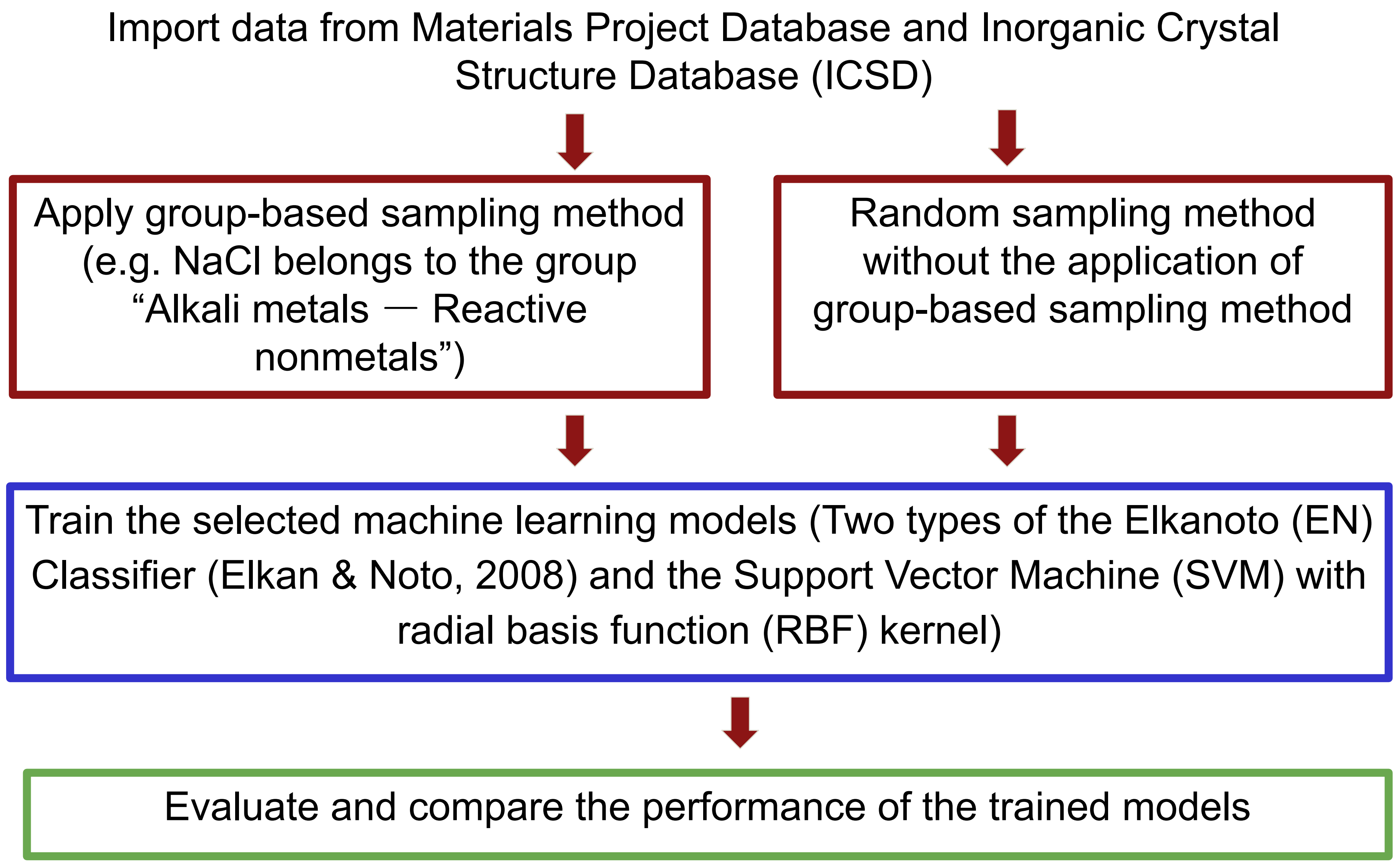
- Development of materials databases have enabled data-driven approaches for searching for new materials (Ward et al., 2016)
- Many data-driven approaches demonstrate the potential to expand the spectrum of synthesizable materials and augment experts' intuitions (Ward et atl., 2016; Cheon et al., 2017; Sendek et al., 2017; Cheon et al., 2018).

Positive-unlabeled (PU) Learning

- PU learning is a subclass of semi-supervised learning and takes a probabilistic approach when dealing with the unlabeled data set.
- PU learning treats the unlabeled data as the superposition of being both positive and negative.



Methods



Import data from Materials Project Database and Inorganic Crystal Structure Database (ICSD)

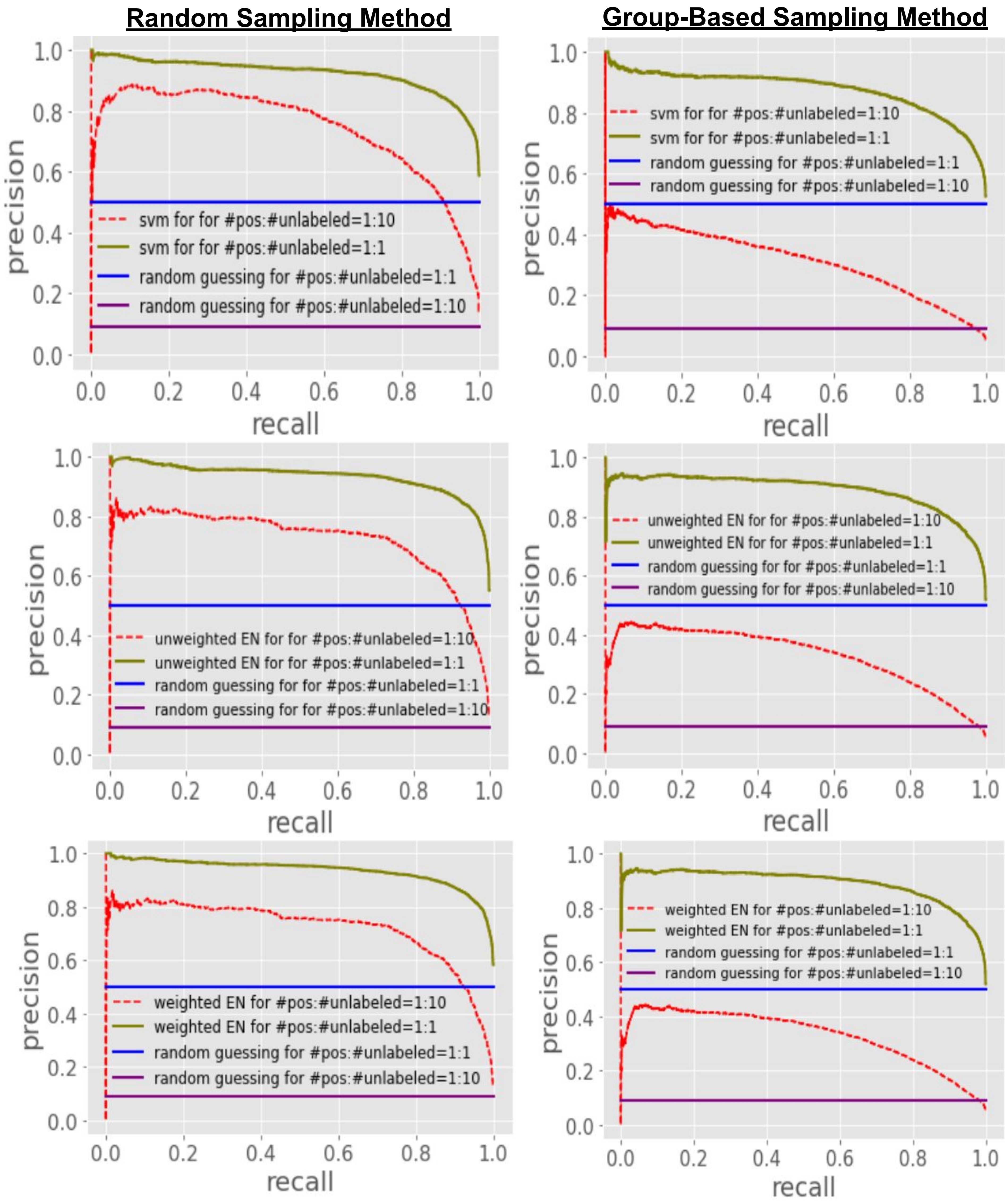
Apply group-based sampling method (e.g. NaCl belongs to the group "Alkali metals — Reactive nonmetals")

Random sampling method without the application of group-based sampling method

Train the selected machine learning models (Two types of the Elkanoto (EN) Classifier (Elkan & Noto, 2008) and the Support Vector Machine (SVM) with radial basis function (RBF) kernel)

Evaluate and compare the performance of the trained models

Results



Random Sampling Method

Group-Based Sampling Method

List of Some Possible Synthesizable Materials

- $Si_2Ir_1Sm_1$
- $Na_1S_2O_2$
- $Si_2U_1Pd_3$
- $Si_1Cu_2Zr_2$
- $Cs_1O_3F_1$
- $Si_3Gd_2Mn_2$

Conclusion

SVM with RBF Kernel

- The SVM has a more noticeable drop in its performance comparing to the EN Classifier when there is a increase in the ratio of the number of unlabeled data to the number of labeled data in the training set.

EN Classifier

- The weighted and unweighted EN Classifiers have nearly the same performance when either the normal random sampling method or the group-based sampling method is applied.

Group-Based Sampling Method

- The drops in the performance of all three models (as the ratio of the number of unlabeled data to the number of labeled data in the training set increases) are more significant when the group-based sampling method is applied.

Stanford | ENGINEERING
Materials Science & Engineering

2020 MatSci REU Symposium. August 28th, 2020.