Band Gap ML part

Anonymous

Paper ID ***

Abstract. Keywords:

1 Introduction

2 Methods

2.1 Machine Learning

In this section we will describe our take on implementing a Machine learning algorithm for Band Gap prediction.

Previous work There were several notable attempts at utilizing Machine Learning to predict physical properties of various materials. Huang et al. reported prediction of band gap properties for ternary metal nitride compounds using ML approach based on calculated hybrid functionals of HSE and DFT PBE and [30]. In their study electronegativity, valence and covalent radius were selected as features for the training of the ML algorithm and prediction. In another study, high accuracy of the prediction was achieved with the ML algorithm trained on a dataset with 3 features of each elements in the compound: ionic radius, electronegativity and number of row associated with position of the element in the periodic table [31].

Data source Firstly, we screened open-access Materials Project and High Performance Computing Center Materials databases to identify stable ternary metal nitride compounds with optical properties that may be suitable for solar energy harvesting. The compounds with predicted band gap values of over 0.5 eV were considered. This is because the DFT method used in the databases for the calculation of the compounds was based on generalized gradient approximation (GGA+U). The latter is known to underestimate the band gap values of the materials relative to the experimentally measured ones [29].

Dataset preparation The algorithm was trained using a dataset of experimentally reported band gap values of about 360 ternary metal nitrides, oxides, sulfides and phosphides compounds.

A lack of data is a common problem in machine learning. This usually imposes limitations on various parts of machine learning pipeline that we will discuss below.

Feature extraction and processing Having low amounts of data also imposes certain limits on the amount of features that can be used without causing the curse of dimensionality [-1]. This suggests us to pick a number of features $d << N$, where $N = 360$ is the total number of data samples available in the training split.

The list of characteristic element-specific features used for the machine learning prediction includes detailed electronic configuration, maximum valence, atomic mas, electronegativity, atomic and covalent radius, ionization potential, electron affinity, period, group and block of the constituent elements.

Model Limited dataset size also affects the choice of the model and limits our choices depending on the model's complexity, since a very sophisticated model, e.g. deep neural network with many parameters, will tend to overfit when the number of training samples is too low.

We used support vector regression (SVR) with the nonlinear radial basis function (RBF) kernel as a machine learning model in our experiments. The model of choice was implemented with the use of the Scikit-learn framework and Python 3.x.

Results The ML prediction of the band gap values was used in combination with the data available in the open access databases to reduce the number of potential candidates for the subsequent theoretical simulations with hybrid functionals of HSE. In addition, the remaining compounds were sorted based on the value of effective electron and hole masses available in the following database [26][27][28].

-1) Trunk, G. V. (July 1979). "A Problem of Dimensionality: A Simple Example". IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (3): 306–307.

References