MATDAT18: Materials and Data Science Hackathon

Team Composition (2 people max.)

Name	Department	Institution	Email
Bharat Medasani	Physical and Computational Sciences Directorate	Pacific Northwest National Laboratory	Bharat.medasani@pnnl.gov

Project Title

High Fidelity Universal Prediction of Bandgaps in Inorganic Materials

Project Synopsis (approx. 100 words)

Electronic bandgap plays a crucial role in material selection for many technologically important applications. Standard density functional theory (DFT) often incorrectly predicts the bandgap and this poses a significant problem for material screening in data-driven material discovery. Post-DFT theories such as hybrid functionals and GW method enable accurate prediction of bandgaps, but are computationally expensive for high-throughput applications**. To remedy this situation, a machine learning based predictor trained on higher fidelity bandgaps with standard DFT bandgaps as one of the inputs is desired for materials across multiple crystal symmetries, chemical bonding types and elemental compositions**. Such a predictor enables quick bandgap prediction for thousands of new materials waiting to be discovered in higher dimensional chemical spaces.

Identified Data-Science Collaborative Need (approx. 100 words)

Existing machine learning studies of bandgaps in materials are limited to a single class of materials with common crystal symmetry, fixed number of elements and composition ratios. Such studies have sufficient data within the narrow chemical space chosen and the input features have uniform shape/size. Our goal of building a universal machine learning model to predict bandgap for different classes of inorganic materials results in a dataset that is sparse for many classes of materials. Further, the number of elements in a structure can vary and this results in different sizes for inputs depending on the class of materials. **An elegant machine learning approach that accounts for the irregular sizes of the input features as well as the data sparsity is required.**

Data Origin and Access (*data must be available and sharable with data science teams* – please address: data source/origin, access privileges, sharing privileges)

The target data is the bandgaps predicted with GLLB-sc functional[4], which is obtained from Computational Materials Repository (https://cmr.fysik.dtu.dk/). Bandgaps predicted with PBE functional (standard DFT) are obtained from The Materials Project (www.materialsproject.org). These datasets are public and can be freely accessed/shared. The input feature data is the elemental properties such as atomic number and compound properties such as crystal symmetry which are either computed automatically or available universally. The input data is curated and preprocessed with scikit-learn.

Project Description (approx. 1.5 pages, plus figures and references; please describe data size, form, dimensionality, uncertainties, number of examples, etc.)

Electronic band structure plays a crucial role in material selection for many applications such as catalysis, photovoltaics, and radiation detection (scintillators). One fundamental and important property of the band structure is bandgap which is the electronic energy range where no states are found and based on which materials are broadly classified into metals, semiconductors and insulators. Often the bandgap is engineered to tailor a material for a target application by varying the composition or by novel nano-synthesis techniques. Mature and accurate modeling tools based on density functional theory (DFT) and post-DFT theories enable avoidance of the costly experimental fabrication route and new materials are often tested *in silico*. Further, new materials can also be predicted *in silico* with the advent of *ab initio* material structure prediction algorithms[1]. These algorithms can explore the ternary, quaternary and higher dimensional chemical spaces that are huge and there are tens or possibly hundreds of thousands of new materials waiting to be discovered in those high dimensional chemical spaces. For *in silico* bandgap engineering to be effective, computationally predicted materials have to be screened with respect to required band structure properties for a target application.

Standard DFT methods often incorrectly predict the bandgap and this poses a significant problem for material screening in data-driven material discovery. Post-DFT theories such as hybrid functionals and GW method enable accurate prediction of bandgaps, but are computationally expensive for high-throughput applications**.** To remedy this situation, a machine learning (ML) based bandgaps predictor trained on higher fidelity bandgaps with standard DFT bandgaps as input is desired. Such a predictor enables quick bandgap prediction for thousands of new materials lurking in higher dimensional chemical spaces.

Recently several independent machine learning (ML) models that use elemental properties as input features have been developed to predict the bandgaps in specific classes of materials[2,3,4]. However each of these models being inherently linear is limited to the specific class of materials from which the training data is sampled. Those models are not transferable to a different class of materials. ML models need careful selection of application and material specific input features, which entails rigorous engineering and in-depth domain knowledge. Redeveloping new ML models for each class of materials will then require huge effort and there could be redundancy in those models.

The materials of interest for any target application could exhibit a wide range of structural and chemical properties. For example, various catalytic materials such as oxides, sulfides, and carbides each have distinct morphologies and chemical bonding. A single bandgap predictor that can work across such different types of materials will make it universal. By combining such predictor with predictors of other properties of interest for a target application, powerful screening tools can be built. Our ultimate objective is to develop ML models with only crystal structure descriptors including elemental composition, and elemental properties as inputs. However, information present in the standard DFT calculations including (incorrect) bandgaps is available at no additional cost after *ab initio* crystal structure prediction step. By utilizing such data, there is a possibility to train more robust ML models.

Post-DFT based evaluation of bandgaps is too expensive even to build a training data set. Instead, bandgaps for nearly 2400 inorganic materials computed with GLLB-sc functional[5] will be used to train ML models. This functional is shown to predict bandgaps that are comparable to those predicted with the costlier post-DFT methods with an uncertainty around 0.2 eV. Majority of the compounds in this dataset contain two to five elements and belong to sixty different crystallographic space groups. The input features to train the model include elemental properties such as atomic number, group number, etc., and compound properties such as lattice constants, crystal symmetry, etc. The dataset has been curated and preprocessed with scikit-learn, which includes transforming the feature space into a zero mean and unit variance data and hot encoding categorical features. After preprocessing, the input data has 165 derived features.

---

[1] A. O. Lyakhov et al., New developments in evolutionary structure prediction algorithm USPEX, Comp. Phys. Comm. **184**, 1172-1182 (2013).

[2] G. Pilania, J.E. Gubernatis, T. Lookman, Multi-fidelity machine learning models for accurate bandgap predictions of solids, Comp. Mater. Sci. **129**, 156 (2017).

[3] J. Lee et al., Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques, Phys. Rev. B **93**, 115104 (2016).

[4] G. Pilania et al., Machine learning bandgaps of double perovskites, Sci. Rep. **6**, 19375 (2016).

[5] I. E. Castelli et al., New Light-Harvesting Materials Using Accurate and Efficient Bandgap Calculations, Adv. Energy Mater. **5**, 1400915 (2015).