# Band Gap Prediction of 2D Materials from 3D Bulk Counterparts

**Disha Dasgupta, Ishaa Bishnoi**
(disha01, ishaa28)@stanford.edu
Stanford University - CS 230

## Abstract

The goal of this research is to predict the band gap of 2D materials through multiple deep learning approaches. We used the easily available properties of the 3D bulk counterparts of 2D materials, such as their bulk band gap, lattice parameters, etc., as the input features. Three models - linear regression, multi-layer perceptron (MLP) and transfer learning - are utilized for this purpose. Among these, the MLP model predicts the band gap with the lowest mean-squared error (MSE) of 0.01. These results demonstrate the power of machine learning to successfully investigate the electronic properties of this emerging class of nanomaterials.

## 1 Introduction

Major technological advancements in nanotechnology are greatly driven by the discovery of new nano-materials. Ever since the discovery of graphene, the study of two dimensional (2D) layered materials has emerged as an important sub-discipline within nanotechnology, electrical engineering and physics. The extraordinary electrical, mechanical, thermal and optical properties of 2D materials, and their successful applications in Field-Effect Transistors (FETs), opto-electronics, sensors, energy devices, plasmonics and spintronics have inspired and renewed interest in exploring more 2D materials.

The band gap (Eg) of a material plays a critical role in determining the electronic and optical properties of the material. Materials with an Eg value of 0 are conductors, while those with some small values are semiconductors. Materials with large Eg values ($> 4$ eV) are non-conducting (insulators). Thus, by knowing the band gap of a material, one can predict its applicability in next-generation devices.

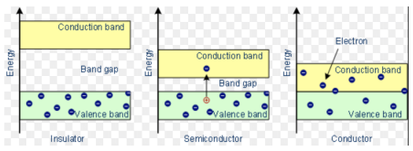


Figure 1: **Semiconductors, insulators, and metals have a small, large and zero band gap respectively [Source: https://www.halbleiter.org/en/]**

The physical properties of nano-materials often change drastically from those of their parent 3D bulk materials. 2D layered materials are one or few atoms thick only, as they are confined in one direction. Therefore, they lack reliable large-scale production methods. Hence, experimentally determining the band gap of a 2D material is challenging, time-consuming and expensive. Thus, computational quantum mechanical modelling methods such as Density Functional Theory (DFT) and Molecular Dynamics (MD) are used to predict the electronic structures and band gaps of 2D materials. However, even these methods, along with the subsequent correction methods such as the GW approach, are very time-consuming and computationally expensive. Since only a few of these 2D materials have been given considerable focus yet and studied extensively, the scientific community is interested in learning about the electronic properties of other unexplored 2D materials, which may be better than the currently studied 2D materials, using quicker and inexpensive methods. Machine learning (ML) has recently attracted attention for estimating the electronic structures of different materials, thus providing a scalable approach that allows one to make informed choices about the materials to focus on for various technological applications.

The objective of this report is to build and optimize various deep learning models that can help provide an overview of the band gaps of the rapidly expanding family of 2D materials.

## 2 Related Work

Most of the research done so far has been limited to finding the band gaps of 3D materials using easily computed properties and information from chemical repositories as features.[8,9] The few works that have

been published on the band gap prediction of 2D materials have focused on specific classes of materials only, such as MXenes and photonic crystals.[1,10] A major challenge with respect to 2D materials is that only a few of them have been discovered or experimentally created so far, and thus, a large database of their properties is currently unavailable. Cheon *et al.* used data mining and identified only 1173 2D layered materials out of over 50,000 inorganic materials in the Materials Project database.[4] No work in our knowledge has been published on predicting the electronic properties of 2D materials using their 3D bulk counterparts' properties. This motivated us to focus our project on predicting the band gap values across all the 2D materials classes using their basic bulk properties, thus creating a broadly applicable model. Although we anticipated high error rates since we used relatively basic properties, our models did decently well as compared to previous works. Moreover, most of these works focused on binary and ternary bulk compounds,[9] whereas we also developed our own model for bulk (3D) band gap prediction as the baseline (pre-trained) model for transfer learning, where we went up to compounds with a maximum of 5 elements. Ferreira *et al.* used MLP to predict the band gaps of 2D photonic crystals, with mean-squared-error (MSE) values of 0.58 and 0.85 for photonic crystals composed of two and three silicon round rods respectively.[10] Similar models for 3D band gap prediction from previous works have reported MSE values around 0.8 for MLP models.

# 3 Dataset and Feature Selection

## 3.1 Dataset

Our models utilize two databases. The first database includes the Perdew-Burke-Ernzerhof (PBE) calculated band gaps of 666 different 2D materials, with band gaps ranging from 0 to 6.5 eV. It has been taken from the Computational 2D Materials Database (C2DB),[2] an openly available database which consists of the structural, thermodynamic, elastic, electronic, magnetic, and optical properties of around 1500 2D materials distributed over more than 30 different crystal structures. Since various compounds occur in multiple structures with different symmetries, we selected only the energetically most stable structure for each compound, hence shortening our database from 1500 to 666 samples. The second database consists of the corresponding 3D bulk properties for each 2D material compound, and is taken from the Materials Project,[3] an open dataset that uses high-throughput computing to estimate the properties of all known inorganic materials.

## 3.2 Features

As can be seen from Figure 2, 585 samples have band gap values between 0 to ∼2 eV. There are 60 individual elements that appear over the full dataset. We also notice that only 4% of these elements occur in 91-110 samples, while 63% of these elements occur only in 1-20 samples.

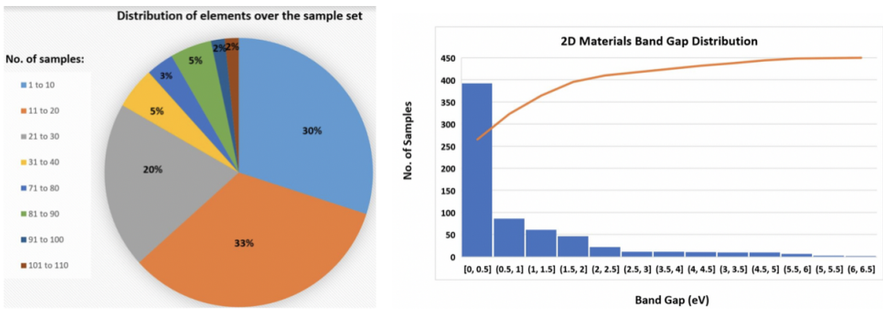


Figure 2: **Distribution of elements over the 2D materials sample set (left) and band gap values distribution in the 2D materials database (right)**

We use a total of 12 input features for each 2D compound which are derived from the compound's chemical formula and its corresponding bulk properties. Since the maximum number of elements present in any compound in our 2D materials dataset is three, we take the number of atoms of the three elements in any sample as 3 of the 12 input features. If a sample has less than three elements, we use a value of 0 for the missing element. For instance, the compound AgBr would use [1 1 0] as its elemental ratio input, $CoPS_3$ would use [1 1 3], and Fe would use [1 0 0]. A more accurate approach would be to use a one-hot vector encoding to represent a sample's elemental composition. This can be done by utilizing a vector of length equal to the number of all the elements present in the entire dataset.[6] Each element in this vector would then represent the number of atoms of a specific chemical element that are present in the sample. For instance, the compound $MoS_2$ will be represented as [1 2 0 0 .....], where the first two elements in the vector represent Mo and S. The MLP model would then be able to learn about each chemical element individually, and consider the presence of each element in a sample accordingly. While we have considered the relative composition in each sample, we did not consider the absolute composition. This is because the dataset size is small, and a few elements like Se and S (which are present in 94 and 110 samples respectively) are dominantly present across the samples, while a few such as Li, Mg, Cs, Si and B appear in only 1-3 samples. This can

cause the model to be under-trained with respect to the less frequent elements in the dataset, producing a high variance. This approach would work well once there are larger databases available for 2D materials properties. In addition, we use the following nine 3D properties taken from the Materials Project:

(1) Lattice Parameters: The six lattice parameters a, b, c, $\alpha$, $\beta$, $\gamma$ of each sample define the physical dimensions (lengths and the angles between them) of its unit cell. (2) Formation Energy (3) Density (4) Bulk band gap

The aforementioned physical properties were selected since they define the basic structure of any material, and simultaneously incorporate the vital physics behind the bonding of the various atoms present in a compound. Figure 3 represents the first few samples in our dataset, along with the input parameters chosen. Element 1, 2 and 3 represent the relative elemental ratio in the sample, as discussed above.

| | Formula | 3D Bulk Band Gap (eV) | 3D Formation Energy (eV) | a (Å) | b (Å) | c (Å) | α (°) | β (°) | γ(°) | 3D Density (g/cc) | 2D Band Gap (eV) | Element1 | Element2 | Element3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | AgBr | 1.171 | -0.372 | 4.476 | 4.476 | 4.476 | 60.000 | 60.000 | 60.000 | 4.92 | 1.724 | 1 | 1 | 0 |
| **1** | AgCl | 0.954 | -0.725 | 3.976 | 3.976 | 3.976 | 60.000 | 60.000 | 60.000 | 5.36 | 1.569 | 1 | 1 | 0 |
| **2** | AgF | 0.000 | -1.194 | 3.556 | 3.556 | 3.556 | 60.000 | 60.000 | 60.000 | 6.62 | 0.497 | 1 | 1 | 0 |
| **3** | AgF2 | 0.000 | -1.455 | 3.711 | 5.128 | 5.051 | 90.000 | 95.392 | 90.000 | 5.06 | 0.000 | 1 | 2 | 0 |
| **4** | AgI | 1.368 | -0.284 | 4.696 | 4.696 | 4.696 | 60.000 | 60.000 | 60.000 | 5.32 | 1.769 | 1 | 1 | 0 |
| **5** | AgI2 | 0.025 | -0.052 | 8.782 | 6.355 | 4.286 | 90.004 | 60.790 | 43.648 | 5.15 | 0.000 | 1 | 2 | 0 |
| **6** | AgO2 | 0.000 | -0.001 | 7.115 | 7.115 | 7.115 | 64.177 | 135.873 | 135.873 | 5.39 | 0.000 | 1 | 2 | 0 |

Figure 3: **First few samples from the 2D materials database with the input features**

The entire dataset was randomly split into 90.25%, 4.75% and 5% training, validation and test sets for the MLP and transfer learning models. Although using smaller values for the validation and test splits gave low errors, the number of samples in these sets turns out to very small since the dataset is small to begin with. Thus, using slightly larger sizes for these sets became necessary to adequately sample across all the materials.

### 3.3 Pre-processing

Many compounds can exist in more than one space group in nature, and each of these exhibit different symmetries in the crystal, as well as different physical properties. To avoid multiple occurrences of the same compound in our dataset, we chose the structure of each compound that is energetically most favorable. Thus, we selected the structure with 0 or the lowest energy above hull, as a zero energy above the hull indicates the most stable material at that particular composition. Furthermore, we limited our dataset to ternary compounds, i.e. compounds with maximum three elements only. The band gap values for the output labels were rounded off to the first decimal place, which is a good enough approximation.

## 4 Methods

We used three different approaches to predict the band gaps of 2D materials: a linear regression model, a multi-layer perceptron model, and a transfer learning approach. Our project employs packages in Python, mainly scikit-learn and keras. The relative elemental ratios are first incorporated into the code to appear as input features in the models.

### 4.1 Linear Regression

For the linear regression model, the data was split 75%, 25% training and test sets based on a general standard for splitting data for prediction models (usually between 80/20 and 70/30). The 2D data was fit to the training set, then the model was used to predict the band gaps in the test set. The MSE was then calculated for the model to determine the model's predictive accuracy.

### 4.2 Multi-Layer Perceptron (MLP)

For the MLP, we first split the 2D dataset into 90.25%, 4.75% and 5% training, validation and test sets. Using this data, we built a three layer MLP model. The dataset is normalized before being available for training. We started off with three layers and decided to change the number of layers accordingly, if needed. The MLP's input, hidden and output layers had 12, 8 and 1 nodes, respectively. The input and output layers were set so as we had 12 input features and 1 output (the 2D band gap). We tested the MLP with different numbers of hidden layer nodes, with the model with 8 nodes in the hidden layer being the most optimal with the best test set MSE. We tested the model with multiple activation functions, such as ReLU and linear functions, for each layer and eventually determined that the model had the optimal MSE when each layer was activated with a Leaky ReLU function. We randomized our inputs with He Normal initialization (based on a general standard - most models are initialized with He or Xavier initialization) and then trained over 300

epochs to avoid over-fitting. We ran our model with an Adaptive Moment Estimation (Adam) algorithm, and used mini-batches of size 32 as these optimized our loss function and computational efficiency. We plotted the loss curve for our MLP to determine the error in the training and validation sets and to ensure we were not over or under-fitting our model. Finally, once our training and validation errors were optimized, we determined the MSE for the test set.

### 4.3 Transfer Learning

Almost all previous works have focused on predicting the band gaps of 3D bulk materials, or various classes of bulk materials (such as organic and inorganic solids), using their physical properties. Therefore, we decided to build a neural network for predicting the bulk band gaps using a 3D materials database of 6105 bulk materials taken from the Materials Project database, and then apply this trained model on our 2D materials database. The input features for this 3D materials database model were chosen to be the following bulk properties of the materials: formation energy, volume, density, number of sites, crystal system, and the relative elemental ratio in each sample. Since the 3D materials database includes samples with a maximum of 5 elements in any given sample material, the model had a total of 10 input features. The crystal systems were encoded into appropriate numerical representations. Given the larger size ( 10 times the 2D materials dataset) of the 3D materials dataset, we decided to transfer the trained weights for the 3D materials model to the 2D materials mode by tuning the last layers of the model to predict the 2D material band gaps.

We first split the 3D materials dataset into 90.25%, 4.75% and 5% training, validation and test sets. The 3D materials database had more samples and data, and thus, the model required more complexity. So we developed a four layer neural network model with the following parameters: the input layer had 6 nodes, the first hidden layer had 4 nodes, the second hidden layer had 2 nodes, and the output layer had 1 node. Initially, we chose our input layer size as 10, given there were 10 input features and 1 output (the 3D band gap). However, after plotting the loss curve for our model, we found our model was over-fitting the data, and thus reduced the number of nodes in each layer.

To apply our model to the 2D database, we froze the first 2 layers so that only that last 2 were trainable. We chose to retrain the last 2 layers as just retraining with a different output layer would not accurately help transfer the model from predicting 3D band gaps to 2D band gaps. We then fit the model on the 2D database to predict the 2D band gaps.

We tested multiple activation functions for each layer, and eventually determined that the model had the optimal MSE when each layer was activated with a Leaky ReLU function. As with the MLP, we randomized our inputs with He Normal initialization (based on a general standard - most models are initialized with He or Xavier initialization), and then trained over 300 epochs to avoid over-fitting. We ran our model with the Adam algorithm, and used mini-batches of size 32 as these again optimized our loss function and computational efficiency.

We plotted the loss curves for both the initially trained 3D model, and the model refitted on the 2D database to determine the error in the training and validation sets, and ensure that we were not over or under-fitting our model. Finally, once our training and validation error were optimized, we determined the MSE for the 2D data test set.

## 5 Results and Discussion

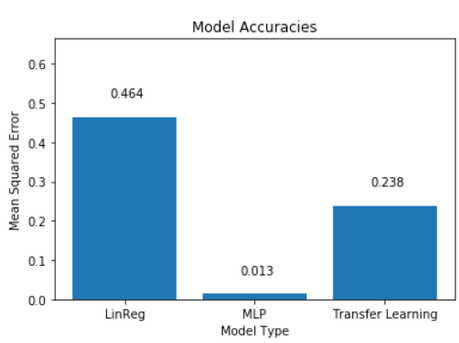


Figure 4: **MSE values for each model**

From Figure 4, we can see that the MLP had the best predictive performance of all the models (lowest MSE of 0.01) and the linear regression model had the worst predictive performance (highest MSE of 0.46).

This is further emphasized by looking at the plots in **Figure 5**, where the actual test set values are plotted versus the predicted test set values for each model.

It is clear from these plots that the MLP's predictions have the highest correlation with the actual test

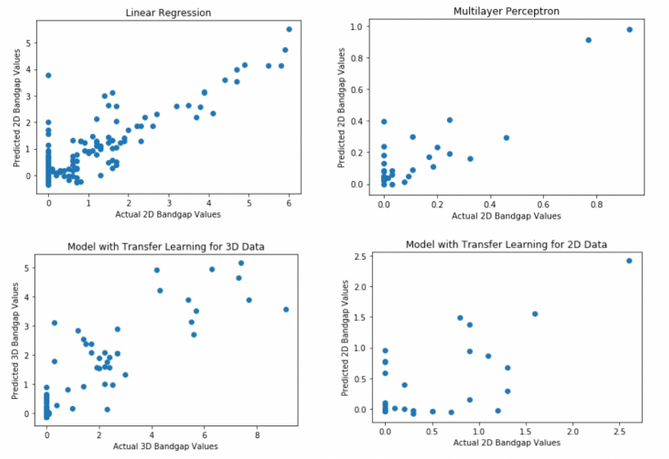


Figure 5: **Comparison of actual vs predicted values for each model**

values. Although the linear regression model may seem to have a higher correlation between actual and predicted values, the linear regression performs poorly on lower band gap values, contributing to a higher test set MSE.

Figure 6 shows the loss curves for the training and validation sets for the MLP, the trained 3D database model, and the transfer learning model for the 2D database. It also indicates that all the three models converge quickly to their minimum losses. As expected, we see these "spikes" in the loss curves as a result of using mini-batches when fitting our models. We see that each of the models are well fitted as the validation set error is not significantly higher than the training set error (which would indicate over-fitting). This can be one of the possible reasons that our models have good predictive accuracy (lower MSEs), especially in comparison to the standards set by previous research.


Figure 6: **Training and Validation loss curves for each model**

# 6 Conclusion and Future Work

Given that the 2D dataset contains only 666 samples, and that basic features were used, the MLP model gives exceptionally good results, with an accuracy that is sufficient to provide useful predictions for 2D materials band gap values. Moreover, better accuracies can be obtained by including more physical, structural and electrical properties of each element present in a sample to increase the complexity of the features, such as their electronegativities, covalent radii, conductivities, melting and boiling temperatures, ionization potentials, group and period numbers in the periodic table, electronic configurations, space groups, and point groups.[5] Pilania *et al.* have shown that the band gap is inversely correlated with the atomization energy, electron affinity and the dielectric constants, and directly correlated with the spring constant.[7] Therefore, including these features would lead to better prediction of band gap values.

We further propose building a Convolutional Neural Network (CNN) for the band structure images for each sample as an alternate method to build a robust model for band gap prediction. While it would be challenging to build such a CNN from scratch, as the images are 1D curves, a pre-trained CNN model that has been trained on similar features can be used.

The next step would be to model a binary classifier that can predict the band gap type (direct or indirect). This would help predict each material's applicability for opto-electronics applications, since some devices such as LEDs and semiconductor lasers require a direct band gap material. We conclude that MLP is a potential approach for band gap prediction of 2D materials, and can be integrated to optimization algorithms to reduce computation time.

## Contributions

Both partners contributed equally to the literature review, conceptual ideas, coding aspects, report writing, and poster design.

## References


[1] Arunkumar Chitteth Rajan, Avanish Mishra, Swanti Satsangi, Rishabh Vaish, Hiroshi Mizuseki, Kwang Ryeol Lee, and Abhishek K. Singh. Machine Learning Assisted Accurate Band Gap Predictions of Functionalized MXene. *Chem. Mater.* 2018, 30, 40314038.

[2] Sten Haastrup, Mikkel Strange, Mohnish Pandey, Thorsten Deilmann, Per S. Schmidt, Nicki F. Hinsche, Morten N. Gjerding, Daniele Torelli, Peter M. Larsen, Anders C. Riis-Jensen, Jakob Gath, Karsten W. Jacobsen, Jens Jørgen Mortensen, Thomas Olsen, Kristian S. Thygesen. The Computational 2D Materials Database: highthroughput modeling and discovery of atomically thin crystals. *2D Materials* 5, 042002 (2018).

[3] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* 1, 011002 (2013); doi: 10.1063/1.4812323.

[4] Gowoon Cheon, Karel-Alexander N. Duerloo, Austin D. Sendek, Chase Porter, Yuan Chen, and Evan J. Reed. Data Mining for New Two and One-Dimensional Weakly Bonded Solids and Lattice-Commensurate Heterostructures. *Nano Lett.* 2017, 17, 19151923.

[5] *Past CS 229 report*: Data-Driven Prediction of Band Gap of Materials. Fariah Hayee, Isha Datye, Rahul Kini. Fall 2016.

[6] *Past CS 229 report*: Predicting electronic properties of materials. Ilan Rosen, Jason Qu and Jacob Mark. Fall 2018.

[7] Ghanshyam Pilania, Chenchen Wang, Xun Jiang, Sanguthevar Rajasekaran Ramamurthy Ramprasad. Accelerating materials property predictions using machine learning. *Scientific Reports* 3, Article number: 2810 (2013).

[8] Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *J. Phys. Chem. Lett.* 2018, 9, 16681673.

[9] Zhang Zhaochun, Peng Ruiwu, Chen Nianyi. Artificial neural network prediction of the band gap and melting point of binary and ternary compound semiconductors. *Materials Science and Engineering* B54 (1998) 149–152.

[10] Adriano da Silva Ferreira, Gilliard Nardel Malheiros Silveira, Hugo Enrique Hernández Figueroa. Multilayer Perceptron Models for Band Diagram Prediction in bi-dimensional Photonic Crystals. *Optics and Photonics Conference (SBFoton IOPC)* 2018 SBFoton International, pp. 1-5, 2018.