Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Graph representational learning for bandgap prediction in varied perovskite crystals

Pravan Omprakash [a], Bharadwaj Manikandan [a], Ankit Sandeep [b,1], Romit Shrivastava [c,1], Viswesh P. [a,1], Devadas Bhat Panemangalore [a,*,2]

[a] *Department of Metallurgical and Materials Engineering, National Institute of Technology Karnataka, Post Srinivasnagar, Surathkal, India*
[b] *Department of Civil Engineering, National Institute of Technology Karnataka, Post Srinivasnagar, Surathkal, India*
[c] *Department of Mechanical Engineering, National Institute of Technology Karnataka, Post Srinivasnagar, Surathkal, India*

ARTICLE INFO

Keywords:
Machine learning
Perovskites
Photovoltaics
Graph Neural Networks

ABSTRACT

Perovskites are an important class of materials that are actively researched for applications in solar cells and other optoelectronic devices due to their ease of fabrication and tuneable bandgaps. High throughput computational techniques like Density Functional Theory (DFT) and Machine Learning (ML) are viable methods to accelerate discovery of new perovskite materials with favourable properties. ML specifically is faster and requires lesser computational power. We recognized the importance of having robust datasets for ML and hence collated a dataset of varied perovskite structures along with their indirect bandgaps. We employed a graph representational learning technique and trained a model that predicted bandgaps for all types of perovskites. The model has a mean absolute error of 0.28 eV and can predict bandgap in a few milliseconds. The metric of generalization gap is introduced to quantify the performance of ML models. This metric will help in building more generalized models that can predict properties for novel materials. Furthermore, we believe that these computational techniques should be user-friendly to those less experienced in the field. Hence, for researchers unacquainted with DFT or ML, we built a pipeline that abstracts the specific processes. This makes it easier for material scientists to quickly screen viable inorganic perovskite compounds allowing them to synthesize and experiment on the more promising compounds.

1. Introduction

Solar energy is considered amongst the prominent sources of renewable energy for the future, to reduce our dependencies on depleting energy resources around us. While the quest to improve this technology is still in pursuit, the introduction of Perovskites based Solar Cells (PSCs) is rapidly taking over the industry because of its immense advantages. Over the past decade, research on PSCs has rapidly increased and a multitude of approaches have been devised to exploit the full potential of the solar cells [1].

Perovskites have the potential for higher efficiency in the conversion of solar energy due to their broad absorption spectrum and tuneable bandgaps [2]. Perovskite structures can be described as a large atomic or molecular cation ‘A’ occupying the body centre of the cell, while ‘B’ cations occupy the corners of the unit cell and smaller ‘X’ anions occupy the face centres. This forms an ABX_3 structure. A perovskite solar cell includes a perovskite structured compound, for example a cesium lead halide, as the light-harvesting active layer. Perovskite materials such as methylammonium lead halides are cheap to produce with relatively simple production methods and low production costs when compared to the large silicon crystals used in conventional solar panels. PSCs can offer higher flexibility, semi-transparency, light-weight, and can have tailored form factors which are characteristics that can allow a plethora of applications of solar cells. Silicon is reaching its practical and economic photovoltaic efficiency limit [3]. Perovskite solar cell technology will allow silicon solar cells to break through the performance barrier and significantly improve the performance of silicon photovoltaics. The power conversion efficiency (PCE) of PSCs has reached a promising value of 25.2% [4], but there is still some scope for improvement. Also, the most promising perovskites are not stable in air and moisture [5]. Hence, better alternatives that are more stable and result in higher efficiency have to be identified.

[*] Corresponding author.
[1] Denotes equal contribution.
[2] 0000-0002-7152-5395

https://doi.org/10.1016/j.commatsci.2021.110530
Received 10 February 2021; Received in revised form 14 April 2021; Accepted 19 April 2021
Available online 6 May 2021

For increasing the efficiency and stability of these solar cells, the perovskite structure space has to be searched for better alternatives. There are millions of possible combinations, and so far only a few thousands have been identified. Thus, computational techniques have to be employed for finding new perovskite materials and their properties in an efficient manner. This trend of using computational methods is expected to continue in the upcoming future because of the amount of data already present. The most popular of these techniques is Density Functional Theory. Density functional theory (DFT) has facilitated a viable substitution to experimental techniques to calculate the properties of materials. DFT is a quantum mechanical modelling method that is used in physics and materials science to investigate the electronic structure of many body systems. It was developed by Pierre Hohenberg and Walter Kohn in the 1980s [6]. It is a highly popular computational technique utilised in materials science to perform high throughput calculations and accelerate the development of materials [7].

Traditional experiments and computational modeling often consume a large amount of time and resources. Furthermore, they are limited by their experimental conditions and theoretical foundations. Efficient and accurate prediction of a diverse set of properties of material systems is possible by employing machine learning methods trained on DFT computations along with the notions of chemical similarity [8]. Machine Learning (ML) for perovskites has been looked into deeply over the past decade for predicting bandgap [9–11], energy over the hull [12], heat of formation [13], stability [14], etc. ML provides an edge over DFT for two reasons. Once trained, the model takes up little space and can predict properties for thousands of compounds in seconds while DFT takes a few hours [15]. Secondly, a trained ML model does not need heavy computational resources (many CPUs, GPUs) to make its predictions, while DFT heavily relies on parallel computation. The short development cycle and the low computational costs make ML a very promising choice for accelerating materials development and in high throughput calculations [16,17]. While the advantages of ML are indeed lucrative, it is important to note that DFT and experimental techniques are the source of data that ML completely relies on.

Over the past few years, graph neural networks have become powerful and practical tools for machine learning tasks in graph domain. They are an upcoming graph representational learning technique now becoming more popular in materials science [12,18,19]. Graph neural networks (GNNs) are deep learning based methods that operate in the graph domain. Due to its convincing performance and high interpretability, GNNs have been a widely applied graph analysis method recently. A massive advantage of using GNNs is that creating complex networks using graph-to-graph building blocks is simpler due to the way GNNs are designed [20]. Furthermore, they offer two advantages over the building blocks of ML: Higher sample efficiency ensuring optimal use of the dataset and its ability to understand and produce novel combinations of already familiar elements which aides the production of combinations of unfamiliar elements.

The MEGNet model makes use of Graph Neural Networks (GNNs) [21] which has acquired a lot of traction in recent times in the field of property prediction in materials. This growing popularity can be attributed to their high feature and representation learning capabilities. The model also has the ability to produce novel results for various problems involving the classification of graph entities and nodes. We utilize this technique to predict perovskite properties in our work.

We believe it is important to use machine learning to help accelerate the development of important materials for critical applications. As mentioned above, Perovskites are an important class for materials for renewable energy, and viable perovskites have to be found rapidly. Finding better materials for photovoltaic technologies will be a step in attaining sustainable energy production. To this effect, we set about to train a ML model in typical fashion.

The foundation of any machine learning model lies on the data. Good, robust data, encompassing a considerable portion of the underlying distribution is of utmost importance for a robust model. The size of the dataset is of paramount importance for better ML models, specifically deep learning techniques. The more the data available, the more accurate the model will be. However, for the model to learn better, the data has to be clean with a uniform distribution. Existing literature has shown that perovskite datasets are scattered, and are of varying sizes and formats. Our effort was to put together as many datasets as we could find in a single, reproducible format. Also, we had to ensure that our dataset would include the different type of perovskites currently under research, such as organometallic [22,23], all-inorganic [24–26], low-symmetry [27], etc.

After collecting a dataset that is a true representative of the underlying distribution, the model can be trained and tested on a hold out dataset to evaluate how well it can generalize to unseen data. Generalizing to unseen data in simple terms would mean making predictions at least as accurate as the training error. We believe that ML models should be built for crystal families to ensure a better understanding of the general structure and properties of that crystal system. A model restricted to a certain specific class of compounds (for example, hybrid inorganic organic perovskites), would not perform well for all inorganic perovskites. This is important as much of material development depends on novel materials that have not been synthesized or modelled before. It is important to at least build machine learning or deep learning models that are not specific to a type of compound. The low, almost negligible errors of the CGCNN [12], GATGNN [21], MEGNet [19] for predicting the properties of inorganic crystals exhibit the ability to carry out this task. However, these models were trained on a substantially large amount of data. As discussed before, a dataset of such a magnitude is hard to collate for perovskites. Although in this paper we have tried to put together such a dataset, it is important to note that further time and effort needs to be given in data production and assimilation either through first principle computational techniques or experimentation.

Furthermore, it is also important that future researchers not familiar with either DFT or ML find it easy to access these methods or for experimental material scientists to be able to use computational resources to quickly shortlist viable compounds for further experimentation. The usability of these processes are also as important as it is to develop them. To this effect, we designed a workflow that enables the experimental researchers or beginners of the field to input various combinations of elements that could form an inorganic perovskite and obtain a bandgap and its crystal structure as an output. Also, this pipeline is made up of discrete units that can be replaced, updated or removed with more efficient or faster solutions. Some of the possible enhancements to our pipeline have been outlined in relevant sections of this paper. We believe that this could be a small step in creating a system that is more user friendly, while also allowing other researchers to develop and enhance it.

The contributions we make in our paper are summarized here:

1) A large dataset has been collated for bandgap prediction of perovskites. A dataset description has also been provided.

2) An existing GNN model has been trained on this dataset and experiments have been performed to interpret the model. Some key features that should be examined for further development of ML models are also highlighted. The model is not trained on specific types of perovskites but on a more general dataset. This effort is to capture general trends in perovskites, and we rationalize that such a model is more important for novel materials development.

3) A pipeline has been created that makes the prediction of bandgaps for inorganic perovskites easier. This facilitates researchers who want to shortlist the compounds for experimental purposes without getting into the specifics of DFT and Machine Learning.

The rest of the paper is organized as follows. A short review on the related works are presented in the Background section. The methods employed to collate the dataset and an analysis of the dataset along with the process used to train the model are presented under the Methods section. The performance of the model is analysed and the development of the pipeline is discussed in the Results and Discussions section followed by the Conclusion which summarizes the paper and provides possible future directions.

2. Background

In this section, the existing literature of various Machine Learning, Deep Learning and graph representational learning for predicting the properties of perovskites are reviewed.

The development of a supervised machine learning approach using support vector classification and support vector regression to predict the bandgaps of inorganic materials without relying on DFT calculated values was reported by Zhuo et al. [28]. Their training set consisted of 3896 experimentally reported bandgaps, and by using experimental bandgap values their predictions were not subject to the same systematic error as DFT determined bandgaps. The resulting model was capable of predicting the bandgap using the elemental properties of the constituent elements, which are related to atom's relative position on the periodic table, the electronic structure, and its physical properties among others. The human intervention for feature selection is a downside to ML techniques, and hence deep learning is chosen which learns to pick the features to enable greater flexibility.

Im et al. [13] presented a machine-learning (ML) based investigation employing a gradient-boosted regression tree (GBRT) algorithm and a dataset of the electronic structures of 540 hypothetical double perovskites of the type $A_2B^{1+}B^{3+}X_6$ calculated using DFT. The algorithm allowed them to obtain accurate models for the prediction of heat of formation and bandgap, along with importance scores for features of the materials enabling them to determine crucial features for prediction and understanding feature-property relationships. A total of 32 features were selected for the model, including chemical information of constituent atoms and geometric information such as bond length and crystal symmetry. With a limited number of structures in the dataset, they obtained an averaged root-mean-square-error (RMSE) of 0.021 eV/atom for test sets of heat of formation which is comparable to the error between DFT-based and experimental values (0.024 eV/atom). In the case of predicted bandgap values, the averaged RMSE of the test sets was found to be 0.223 eV which despite being less accurate is an acceptable error considering the effective range (1.1–1.8 eV) suitable for solar cell applications. There was manual feature selection which is a major disadvantage of ML techniques. Moreover, for the dependable training and evaluation of ML models, it is important to have larger dataset sizes.

A supervised machine learning model using the kernel ridge regression method to predict electronic, geometric and thermodynamic properties of perovskites was devised by Stanley et al. [29]. The model was trained using a new database of 344 mixed perovskites created using DFT with Cs^+, Rb^+, K^+ or Na^+ being the possible ions at the A site, Sn^{2+} or Ge^{2+} at the B site and I^-, Br^- or Cl^- at the X site. They achieved an average root mean square error of 146 ± 19 meV when applying it to around 100 random test-train splits of 10% and 90% respectively of the dataset. This work utilized mixed perovskites, which is now a highly researched type of perovskites. However, they note that larger datasets are required to gain higher confidence in trend and property prediction.

Gladkikh et al. [30] presented multiple machine learning approaches that describe the correlation between the bandgap of perovskites and their constituent ions' elemental properties. The non-linear mappings between bandgap and elemental properties were learned using alternating conditional expectations (ACE) a semi-parametric machine learning technique suitable for smaller datasets. They concluded that the bandgap is mostly determined by electronegativities, electron affinities, ionisation energies, and atomic radii of the constituent ions. The different machine learning models' performance was compared using the root mean square error (RMSE) and mean absolute error (MAE) and concluded that kernel ridge regression achieved the best performance.

An implementation of a random forest classification model to predict the bandgap of perovskite materials with the number of trees set to 10 was reported by Takahashi et al. [31]. The dataset used for their work consisted of 15000 perovskites generated by the first-principle calculations. The accuracy was cross-validated by randomly splitting the dataset into 80% trained data and 20% test data, with the average score being 0.98 with a standard deviation of 0.002 for ten randomly split trained and test data. They then predicted 9328 perovskites with a bandgap in the range 1.7–3.0 eV; these 9328 perovskites were further investigated for their thermal stability and bandgap using first-principle calculations. They concluded that there were ten thermodynamically stable undiscovered Li and Na based perovskites with ideal bandgaps for capturing solar light. It is to be noted that a classification approach was taken to predict whether a perovskite had a bandgap in a certain range. However, the use of a large dataset for the purpose is a highlight of this paper.

A predictive model for the electronic properties of metal halide perovskites (MHPs) was developed by Saidi et al. [32] using Convolutional Neural Networks (CNN).A dataset including structural and electronic properties of 862 MHPs was systematically prepared. The perovskites used were mainly of the form $MAPbI_3$ lattice, owing to their 25.2% power efficiency. A computational screening was performed on ABX_3 MHPs, focusing mainly on variations at the "A-ion" site, while keeping the rest of the lattice similar to the $MAPbI_3$ structure. A Hierarchial Convolutional Neural Network (HCNN) was used for the task. The model exhibited promising root-mean-square errors (RMSEs) for the lattice constants, octahedral angle and bandgap. They were 0.001 nm, 5°, and 0.02 eV, respectively.

Zlatomir Stoichev et al. [33] makes use of artificial neural networks for the analysis and prediction of 3D perovskite bandgaps. The set of input parameters which were used for the artificial neural network included parameters such as s, p, d, and f orbital radii, electronegativity, octahedral factor, and formation energy. A recursive correlation filter was used to remove features that had high correlation with their model. Along with this, the features with low contribution to the bandgap were eliminated too. The model yielded a mean square error approximately equal to 0.25 eV. The deep learning techniques while increasingly robust, do not exploit the correlation between atoms and their properties. Graph representational learning techniques seek to do this.

The generalised Crystal Graph Convolution Neural Network (CGCNN) framework was presented by Xie et al. [12] to represent periodic crystal systems and provide material property prediction achieving comparable accuracy with respect to density functional theory (DFT) calculations as DFT compared with experimental data. Their approach utilised a CNN built over a crystal graph that represents the structures by encoding both atomic information and bonding interactions, which enabled them to extract representations that are optimal for predicting properties by training with DFT calculated data. The generality of the framework was investigated by observing prediction performance for different material properties, namely, absolute energy, bandgap, shear moduli, Poisson ratio, etc. It was found that the mean absolute errors (MAEs) were close to or higher than DFT accuracy when training data of the order 10^4 were used. They further used the CGCNN framework to predict the total energy above the hull of perovskites and achieved an MAE of 0.130 eV/atom on 3787 test perovskites. However, they did not take into account the universal features of the system (for example temperature).

Louis et al. [21] proposed a deep graph neural network named GATGNN based on a global attention mechanism that was first used in neural networks for natural language processing. Their model used local attention layers to capture local atomic environments' properties and a global attention layer to create the global representation of the whole crystal system by making weighted aggregates of all the atomic environments. The model's prediction performance was tested and compared with the results of the CGCNN and MEGNET models. They found that their model showed a 10% score improvement over CGCNN for absolute energy, bandgap and bulk moduli. It showed improved prediction for bandgap, bulk moduli and shear moduli whereas it did not perform as well in the prediction of heat of formation when compared to the MEGNet model. MEGNet also had the concept of elemental embeddings which were pivotal for transfer learning.

It can be seen that there have been numerous approaches and ML techniques that have been employed to predict the various properties of perovskites including bandgap. However, the ML techniques require human intervention for the selection of features. The dataset size is also important for dependable and less random results. The DL approaches while learning the features themselves, do not exploit the crystal structure and the global properties effectively. To learn this important information, graph neural networks were built that performed extremely well for crystals. MEGNet provide useful tools for transfer learning to relatively smaller dataset sizes and hence it is useful for our purpose. Lastly, all the models focus on a certain type of perovskite, while we endeavour to build a model encompassing as many types of perovskites as possible.

3. Methods

3.1. Dataset collation

Upon reviewing the recent literature and collecting nine data-sets in all - seven from the Computational Materials Repository (CMR) [36], and two other datasets compiled by Kim et al. [23,35], it was decided that all these datasets must be compiled and stored in a common, reproducible format to facilitate the ease of access and analysis.

The nine datasets mentioned above were cleaned using various libraries in Python, a notable one being Pymatgen (Python Material Genomics) [37]. The duplicate perovskites were eliminated following which all the datasets with varying features were condensed into a large, common dataset with the cell structure parameters and bandgap (position of atoms, lattice parameters, symmetry). These features helped to convert the dataset into 27400 CIF (Crystallographic Information Framework) files. The reason for converting to a CIF format was considering their superiority of representing material information [38].

This led to the creation of a unified perovskite dataset for the bandgap values, which will be released to help further computational research on perovskites. The description of the individual datasets are provided in Table 1. Castelli et al. [24], used DFT calculations to compile a dataset of 19000 oxides, oxynitrides, oxysulfides, oxyfluorides, and oxyfluoronitrides in the cubic perovskite structure targeting PEC applications. Castelli et al. [27] further combined 3D oxide and oxynitride perovskites to investigate 300 layered perovskites in the Ruddleson-Popper phase [39]. Castelli et al. [34] investigated the band gaps and optical spectra of functional perovskites composed of layers of the two cubic perovskite semiconductors; $BaSnO_3$-$BaTaN_2O$ and $LaAlO_3$-$LaTiN_2O$. Organometal perovskites were investigated for their bandgaps [22,26]. Kuhar et al. [25] screened perovskites of the type ABS_3 and used their findings to synthesize promising perovskites for water splitting. Kim et al. [23,35] created a dataset of over 3000 hybrid organic–inorganic perovskites, and also screened thousands of compounds and predicted the dielectric breakdown strength of 200 compounds with a machine learning model.

Table 1
Dataset statistics. A – Cation, A* – Organic cation, B – Cation, B* – Inorganic cation, M – Metal cation, S – Sulphur, X – Halogen.

Dataset	# samples	Type of Perovskite	Reference
Dataset of ABS_3 perovskites	4079	ABS_3	Kuhar et al. [25]
Organometal Halide Perovskites	240	A^*MX_3	Castelli et al. [22]
Perovskite water-splitting database	19369	ABX_3	Castelli et al. [24]
Absorption spectra of Perovskites	79	ABX_3	Castelli et al. [26]
Low symmetry perovskites	1984	$(ABX_3)_n$	Castelli et al. [27]
Functional perovskites	94	ABX_3	Castelli et al. [34]
hybrid organic–inorganic perovskites	1346	$A^*B^*X_3$	Kim et al. [23]
ML assisted perovskites prediction	209	ABX_3	Kim et al. [35]

In total, the combined dataset has 27400 perovskites of varied crystallography and nature which were converted into 27400 valid CIF (Crystallographic Information Framework) files. Two examples of the compounds taken from the dataset are shown in Fig. 1. The dataset contains perovskites of different types such as double perovskites, 2D perovskites, organometal, all-inorganic etc. This is very important when a model is trained on this complete dataset. It would be more robust in predicting bandgaps for any of these above mentioned types. However, it is also imperative that more data is added on the other types of perovskites. We believe it is an important future direction.

3.2. Dataset analysis

An analysis of the dataset was also performed to understand the shortcomings of current major perovskite datasets and also capture trends in the data distribution. The emphasis was on analysing the most frequently occurring elements in the dataset. Lead halide perovskites are the most popular options for solar cells due to their promising efficiencies [40]. However, stability issues of lead halide perovskites hinder their progress and many other substitutes have been explored in the past decade [41]. Lead toxicity in PSCs are a major concern for health and environment issues [42]. This trend is seen in the data with 308 lead halide based perovskites and 1418 lead-based perovskites.

Overall, 64 different elements have been explored in various combinations. The number of occurrences of the A site elements and B site elements are shown in Fig. 2. Halogens are still the most popular choice for anions in perovskites with over 9300 samples containing halogens. Oxygen, sulphur, oxynitride and oxyfluorides are other commonly investigated elements in the X site, due to the efforts of Castelli et al. [24] and Kumar et al. [25]. Table 2 lists the most frequently used organic cations in the dataset.

With respect to the bandgap data collected, one major realization was that there were 19089 samples with an indirect bandgap of 0 eV indicating the presence of 19089 metals in the dataset. This forms a significant proportion (70 %) of the entire dataset, which is undesirable for any machine learning model to train on. This imbalance of data is a major shortcoming for the current perovskite datasets for solar cell purposes. It is to be noted that this imbalance between metallic perovskites and non-metallic perovskites was the main problem considered when choosing our model. Additionally, there are 2879 null values, indicating the presence of 2879 perovskites whose bandgap values do not exist in the dataset and have hence been labelled null given they are not known. This reduces the dataset to 24501 meaningful samples that can be utilised to train a neural network. Also, the number of perovskites having bandgap values in a desirable range of 1–4 eV for semiconductor applications are 3624 (as shown in Fig. 3).

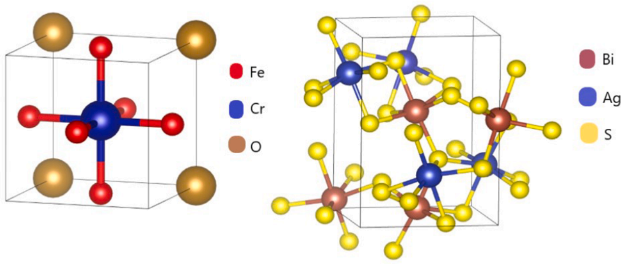


Fig. 1. a) Unit cell of the cubic perovskite; $CrFeO_3$, b) Unit cell of ABS_3 type perovskite; $BiAgS_3$.

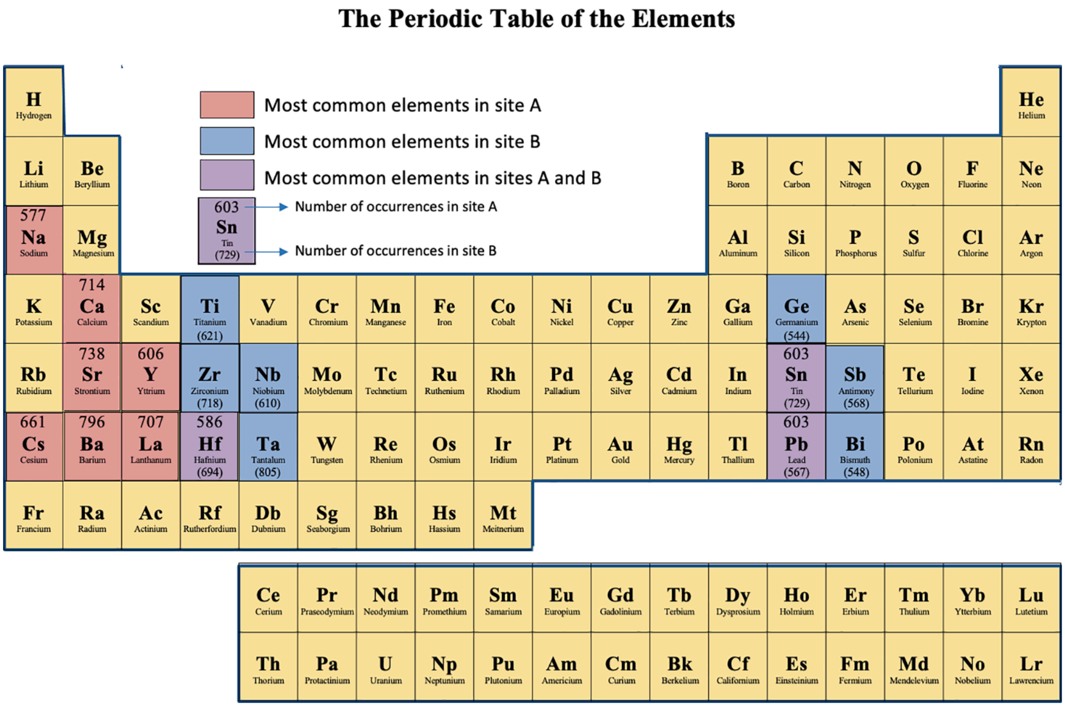


Fig. 2. The most commonly occurring elements in the cation sites.

Table 2
Most common organic cations used in the A site [23]

Organic Cation	Number of occurrences
Azetidinium ($C_3H_8N^+$)	171
Propylammonium ($C_3H_{10}N^+$)	137
Ethylammonium ($C_2H_8N^+$)	131
Hydrazinium ($N_2H^+{}_5$)	116
Isopropylammonium ($C_3H_{10}N^+$)	115
Trimethylammonium (C_3H_9N)	113
Dimethylammonium ((C_2H_8N)	87

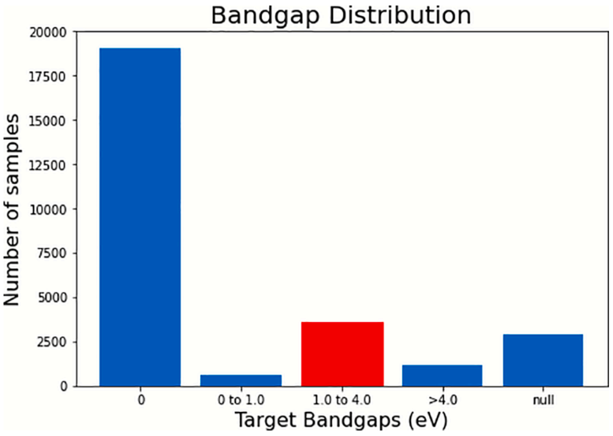


Fig. 3. Bandgap distribution. The range of bandgaps from 1–4 eV has been highlighted in red to underscore the importance of adding more semiconductor perovskite data.

More data about perovskites having bandgaps in the semiconductor region have to be added to this dataset. ML and DL models require a more complete and uniform data distribution to make accurate predictions, and this is an important direction that needs to be considered for using Data-assisted techniques for characterization of perovskites. The dataset have been released on our github repository (https://github.com/Pravanop/Perovskite-Prediction).

3.3. Model used

The model that we chose to use for predicting perovskite bandgaps is the MEGNet (MatErials Graph Network) framework [19,43].

The Materials Graph Network is an implementation of DeepMind's graph networks for universal machine learning in materials science [20]. Its success in achieving very low prediction errors in a broad array of properties in both molecules and crystals has been demonstrated before [19]. Further reasons for choosing MEGNet over other graph learning models have been discussed below.

Given the size of the dataset used for training (10,000), MEGNet was an optimal option as it has been demonstrated to exhibit high rates of improvement for similar large sized datasets as the model can leverage this information more efficiently [44].

While GATGNN [21] has shown promising and comparable results to MEGNet, one of the hallmark features of the MEGNet model was the inclusion of elemental embeddings, which was missing in GATGNNs. Additionally, the MEGNet model was easier to implement and the process was made simpler because of continuous maintenance by the developers. MEGNet allows for a robust foundation for building models that contain state properties, making it more generalizable [19].

The elemental embeddings in the MEGNet model [19] encode periodic chemical trends which aid the process of transfer learning that can consequently be used to predict properties with limited training data. Thus, a shortage of data regarding interpretable chemical trends can be solved by extracting the required values from the elemental embeddings trained on a large data set. This helps improve the performance of models with smaller data quantities. This model has three types of attributes [19] as inputs - atomic attributes, E (atom type, chirality, hybridisation, etc.), bond attributes, V (bond type, bond order, graph distance, etc.) and global state attributes, u (temperature, average atomic weight, bonds per atom, etc.).

3.4. Model training

As already mentioned before, the dataset consisted of many metallic compounds and consequently results a bandgap value of zero. This caused a severe imbalance of data causing the model to not learning anything new. Studies have been made to improve the imbalance of datasets in a regression setting [45], however, the results are not compelling enough to try for this dataset. Hence, 85% of the metals in the dataset were removed, that has considerably decreased the size of the dataset. The reason for removing 85% of the metals was to approximately equalize the number of 0s and no zero values. This results is much more uniform distribution than before. Therefore, 8001 values remained, that were used to train the model. The data was randomly split into 2 parts, 1) Training, 2) Testing in a ratio of 6:1 and proceeded to train it using our model.

The metric error used was the Mean Absolute Error (MAE), i.e, the arithmetic average of the absolute error between predicted values y_i and true values x_i, for n number of samples. It is calculated as follows:

$$\mathbf{MAE} = (\frac{1}{n})\sum_{i=1}^{n}|y_i - x_i| \tag{1}$$

The hyperparameters were chosen on a trial and error basis. A learning rate of 0.00075 was used, with a batch size of 32. A lower l2 regularization coefficient of 0.001 was chosen, to ensure better generalizability of the model [46]. Two more hyperparameters that are specific to the MEGNet model; elemental embeddings and number of blocks were studied further.

As already mentioned before, the dataset consisted of many metallic compounds and consequently results a bandgap value of zero. This caused a severe imbalance of data causing the model to not learning anything new. Studies have been made to improve the imbalance of datasets in a regression setting [45], however, the results are not compelling enough to try for this dataset. Hence, 85% of the metals in the dataset were removed, that has considerably decreased the size of the dataset. The reason for removing 85% of the metals was to approximately equalize the number of 0s and no zero values. This results is much more uniform distribution than before. Therefore, 8001 values remained, that were used to train the model. The data was randomly split into 2 parts, 1) Training, 2) Testing in a ratio of 6:1 and proceeded to train it using our model.

The metric error used was the Mean Absolute Error (MAE), i.e, the arithmetic average of the absolute error between predicted values y_i and true values x_i, for n number of samples. It is calculated as follows:

$$\mathbf{MAE} = (\frac{1}{n})\sum_{i=1}^{n}|y_i - x_i| \tag{2}$$

The hyperparameters were chosen on a trial and error basis. A learning rate of 0.00075 was used, with a batch size of 32. A lower l2 regularization coefficient of 0.001 was chosen, to ensure better generalizability of the model [46]. Two more hyperparameters that are specific to the MEGNet model; elemental embeddings and number of blocks were studied further.

3.5. Elemental embeddings and number of blocks

To test the robustness of these embeddings, we train models with and without elemental embeddings (called as M(E) and M respectively). The experiments were run for 100 epochs, as there was only a need to make comparative analysis. The experiments were run three times and the average error of the three were taken. M(E) performed well, which was as expected. For 100 epochs, M(E) obtained a training error of 0.38 eV, and testing error of 0.40 eV, while M obtained a training error of 0.42 eV, and a testing error of 0.48 eV. Moreover, the training time required for M(E) was lesser by 30%. These results indicate that the inclusion of elemental embeddings certainly helps the model in achieving lower errors. Also, the generalizability of M(E) is better. Generalizability of a model can be quantified by a generalization gap (G). G can be defined as the absolute difference between training and testing errors. M(E) evidently has a smaller G value.

Next, the number of typical MEGNet units (as shown in Fig. 4) was varied from 2 to 5, and the model performance was checked. 5 blocks of a typical MEGNet unit was chosen as the upper limit, as a higher number of blocks will obscure the contribution from elemental embeddings (as stated in [19]). The results of varying the number of blocks are formulated in Table 3. Increasing the number of blocks facilitates more information flow in the graph, aggregating important information from more nodes farther off in the graph. As seen from Table 3, 5 units gave the best performance on both training and test datasets. This implies for learning perovskite bandgap, information from many unit cells were needed. This can be attributed also to the large number of different perovskites (organometallic, metal halide, all inorganic, 2D perovskites, double perovskites, etc) present, and to make a generalized function that describes the underlying distribution, larger amount of information from farther nodes had to be collected. We trained one model that encompasses all the various types, because we believe that future work should work on building models that are universal for a certain crystal system, with the practical limitations of data and computational resources.

According to the results in Table 3, the model with 5 blocks, and with elemental embeddings are chosen and the model was trained for 1000 epochs, making use of the hyperparameters reported in the previous section. It took 19772 s to train on a single 12 GB NVIDIA Tesla K80 GPU. The training error was found to be 0.245 eV.

4. Results and discussions

4.1. Model performance

The trained model was tested on a hold-out dataset of 1201 compounds with no overlap with the training dataset. The model obtained a MAE of 0.28 eV. This is comparable with the accuracy of DFT calculations [25]. Moreover, given the crystal structure, it took only 89 s to predict 1200 compounds on a single 12 GB NVIDIA Tesla K80 GPU. This means that only 0.007 s was needed for our model to predict bandgap for a single compound, with a GPU. Even if there are further computational contraints, the time needed would still be modest. DFT would require far more time to compute the indirect bandgap of a crystal. As discussed before, this is the major advantage that make machine learning models favourable than other first principle computational methods.

To evaluate the model's performance, apart from the MAE, some other important graphs and metrics were calculated. The predicted bandgaps are plotted against the actual bandgaps (as shown in Fig. 5b), to exhibit the low variance of the predicted bandgaps with respect to the calculated bandgaps. The red points in the graph indicate the predicted bandgaps of those perovskites that have their actual bandgap in the semiconductor range of 1.0 to 4.0 eV. The black line is the x = y line. The lesser is the error of the model, the closer the points should lie to this line. This would mean that the predicted targets are almost equal to the actual targets. Fig. 5a) is the distribution of errors across the 1200 samples. It can be seen that 536 samples have an error below 0.1 eV. This accounts for nearly 45% of the test data. Also, 50% of the data has an error below 0.2 eV. The average of this distribution, or in other words, the MAE, is close to 0.30 eV. 804 test samples have an error below 0.30 eV, which is around 67% of the data. This plot helps in quantifying our model performance in yet another way. We can say that 67% of the time, our model will predict a bandgap with an error lesser than 0.30 eV.

To ensure that the model can predict metallic perovskites reliably, we tested the model on the zeros that we removed earlier. The MAE of the model on only these metallic perovskites was 0.1 eV. The model can

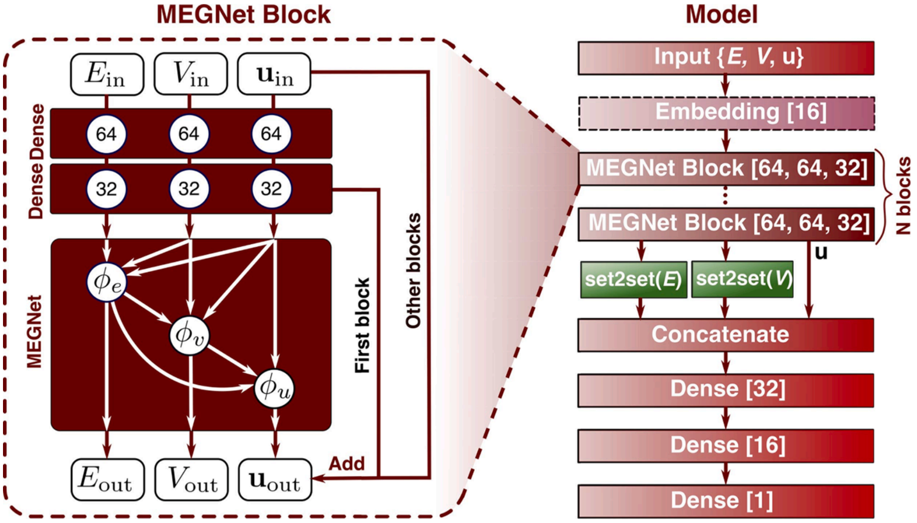


Fig. 4. Schematic diagram of a typical MEGNet structure formed by stacking multiple MEGNet blocks. The three attributes - atomic, bond and global state are inputted into the dense layers. This is followed by vector reduction and concatenation. The MEGNet module contains a series of update operations that maps it from an initial state of G = (E, V, u) to G* = (E*, V*, u*). Reprinted with permissions from Chen et al. [19].

Table 3
Performance of different number of MEGNet units trained for 100 epochs

Number of MEGNet units	Training loss	Testing loss
2	0.45	0.41
3	0.44	0.45
4	0.42	0.39
5	0.40	0.38
6	0.42	0.40

hence classify metallic and semiconductor perovskites with good accuracy.

The model performance was quantified further by establishing the R-squared value. R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable (V_p) that is explained by an independent variable or variables (V_a) in a regression model as shown in Eq. (3). It provides a measure of the ability of the model to correctly predict the variance of the true distribution. The R^2 value of 1 indicates that the model can accurately predict the entire distribution of the true values. Our model obtained an R^2 value of 0.91. An R^2 value of 0.91 would imply that the model explains 91% of the variation witnessed in the target variable, i.e., predicted bandgaps.

$$\mathbf{R}^2 = 1 - \frac{V_p}{V_a} \tag{3}$$

4.2. Pipeline

As shown in Fig. 6, a process was created to allow for the ease of prediction of indirect bandgap of single or double inorganic perovskites in cubic or orthorhombic spacegroups. The pipeline takes in the elements at different sites of the perovskites with their ionic radii and outputs the corresponding bandgap and a CIF file of the structure, if its a valid compound. The individual components of this pipeline are described below.

Tolerance compute. A necessity to characterize materials and their properties is the knowledge of its crystal structure. While there are many methods, the Goldschmidt Tolerance Factor is one that has played a pivotal role in the development of solid state perovskites for years [47]. This has been extended to various classes of perovskites like organic–inorganic, halide, 2D perovskites, double perovsktes etc., and exhibits a strong agreement with the experimental results.

Its utility does not end at finding the crystal structure alone. It has been used with great success to understand the geometric stability, formation, mechanistic origins [48], ion compatibility, strain, charge separation amongst various other properties [47]. However, here it is only used as an initial screening factor for stability in a given crystal structure.

There have been concerns raised about the viability of this factor in recent years. According to a study performed on 576 experimentally characterized compounds, tolerance factors within the range of 0.825 to 1.059 yielded a middling accuracy of 74% [49]. Additionally, there have

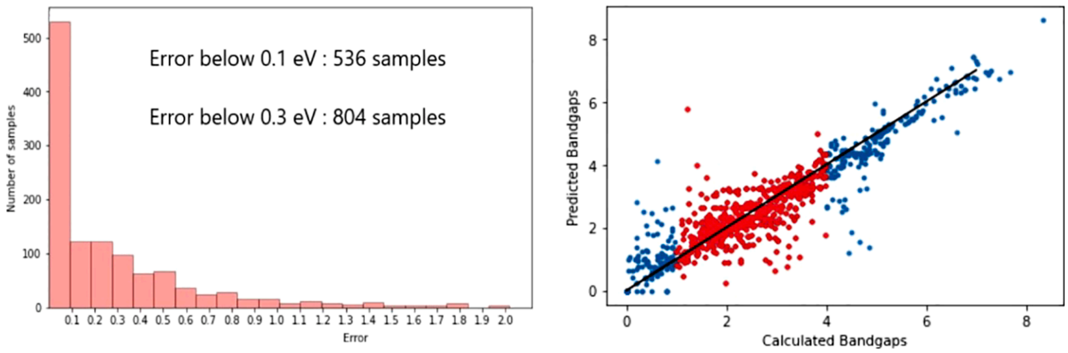


Fig. 5. a) The distribution of Mean Absolute Errors b) The actual targets vs the predicted targets in eV.

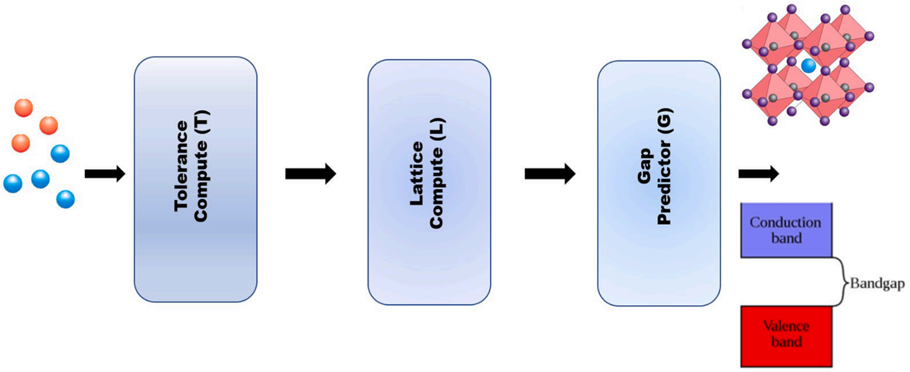


Fig. 6. Schematic diagram of pipeline.

been issues raised regarding its accuracy given the dependence of ionic radii on temperature. Stability predictions for perovskites using ML has been developed [14], and if there are more promising results, these models could possibly replace the Goldschmidt factor.

However, the clear edge of using the Goldschmidt Tolerance Factor (T) is the simplicity in method. In perovskite compounds with the general formula ABX_3 (where A and B represent metal cations and X represent elements from the chalcogen and halogen groups), the tolerance formula can be calculated using a simple ratio [50] of the constituent ionic radii as shown in Eq. (4):

$$\mathbf{T} = \frac{R_A + R_X}{\sqrt{2}(R_B + R_X)} \tag{4}$$

In the equation above, R_A, R_B and R_X are the ionic radii of A, B and X respectively. The structures shown by perovskites according to the Goldschmidt tolerance factor are summarized in Table 4.

Goldschmidt tolerance factor still plays a vital role for analysing double perovskites and other mixed perovskites [51]. The equation is changed to account for the average ionic radius in the site as shown in Eq. (5):

$$\mathbf{T}^{'} = \frac{R_A^* + R_X}{\sqrt{2}(R_B^* + R_X)} \tag{5}$$

For a double perovskite $AA^*BB^*X_6$, R^*_A is the average of the two radii, A and A*,while R^*_B is the average of the two radii, B and B* respectively.

Lattice Compute. To compute the lattice constants for either cubic or orthorhombic structure, DFT in the Projected Augmented Wave (PAW) formalism is used [52]. All calculations were made with a GPAW calculator supported by the ASE environment [53]. The cell relaxations are made using the PBE functional [54]. The wavefunctions are expanded in a plane-wave basis with an energy cutoff of 500 eV. Monkhorst–Pack scheme [55] was used for sampling the Brillouin zone. For cubic structure a 8*8*8 k-point mesh is used, while for orthorhombic structure a 6*6*6 k-point mesh is employed. DFT has been used widely as a first principle method to compute cell structures. Deep learning techniques [56] have been employed to predict the lattice constants of perovskite structures, and to speed up the process these can be considered as viable options to fully remove DFT calculations from this pipeline.

GapPredictor. We employ our trained MEGNet model to finally predict the indirect bandgap (Eg) in eV, and we use pymatgen [57] to convert the structure into a cif format file for later reference.

Table 4
Tolerance factor correlation perovskite structures

Goldschmidt Tolerance Factor (T)	Structure	Example
>1	Hexagonal or Tetragonal	$BaNiO_3$
0.9–1	Cubic	$SrTiO_3$
0.71–0.9	Orthorhombic/Rhombohedral	$CaTiO_3$
<0.71	Different Structures	$FeTiO_3$

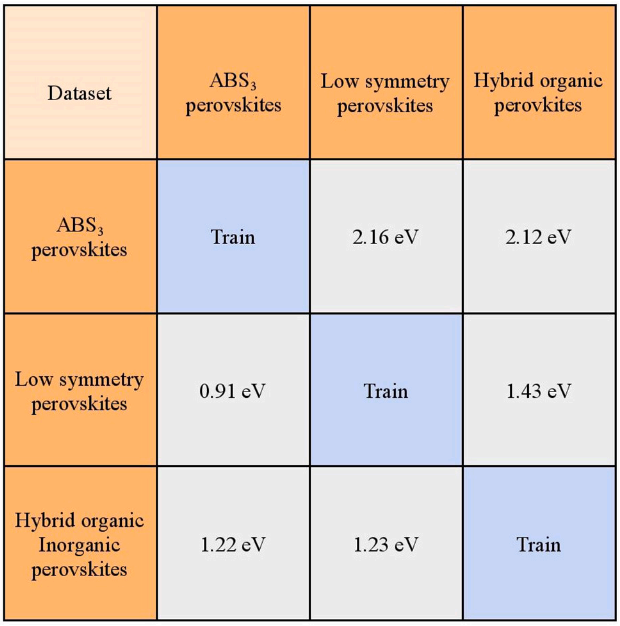
The pipeline accepts only inorganic perovskites currently, but can accomodate organometal perovskites as well as our model is trained on these compounds and can make reliable predictions. We have released this pipeline on our github repository (https://github.com/Pravanop/Perovskite-Prediction) to provide access to any researcher unacquainted to the techniques employed in computational materials science, so that they can predict bandgaps of single or double inorganic perovskites.

4.3. Generalizability of the model

To further understand if the model learnt about the underlying data distribution, i.e., of any general perovskite (orga-nometal, double, inorganic, low symmetry etc.), an experiment was performed by fixing the hyperparameters (learning rate, regularization coefficient, number of blocks, etc.) but varying the training and validation datasets. Usually the entire dataset is split randomly into training and validation sets and performance of the model is reviewed. This is called as K-fold validation. However, we took a slightly different approach, by making the splits in a non-random manner.

We picked out the ABS_3 type [25], low symmetry perovskite [27] and the hybrid organic–inorganic perovskite compounds [23] from the compiled dataset. The reason for picking these three 'sub-datasets' is due to the singular reason that the number of samples were in the order of 10^3, and a ML model's performance is reliable only when the experiments are performed with a considerable sample size. We cannot underscore the importance of having more data for both training and testing. Small testing datasets cannot be used to evaluate the performance of a model. Larger datasets will reduce randomness while also being a good enough representative of the data distribution. These 3 sub-datasets also had good distribution of bandgaps. The model was trained on one dataset and tested on the other 2 datasets, and this was done iteratively. Fig. 7 shows the resulting generalization gap of all the combinations. As seen under the "Low symmetry perovskites" column in Fig. 7, the model when trained exclusively on low symmetry gave an error of 2.16 eV and 1.23 eV when tested on ABS3 perovskites and Hybrid Organic Inorganic Perovskites sub-datasets respectively. The average generalization gap was calculated, and was found to be 1.51 eV. This would mean that on an average, if the model is trained on a certain subset of perovskites, it would incur an error of 1.51 eV when tested on other subsets of perovskites, not including the subset it was trained on.

This generalization gap, G can be quantitatively represented as the difference between the training accuracy and testing accuracy on separate datasets and a high G value indicates the poor performance of the model on unseen data. However, in our case it can be easily seen that the sub-datasets have different perovskite structures, and lesser number of samples for the model to fully learn a general function that can predict accurate bandgap values for any given structure. It is expected of the



Dataset	ABS_3 perovskites	Low symmetry perovskites	Hybrid organic perovkites
ABS_3 perovskites	Train	2.16 eV	2.12 eV
Low symmetry perovskites	0.91 eV	Train	1.43 eV
Hybrid organic Inorganic perovskites	1.22 eV	1.23 eV	Train

Fig. 7. Generalization gap of the various validation tests. The Y-axis here is dataset we tested our model on and X-axis is the dataset we trained our model on..

model to generalize poorly in these conditions, and only with more data can this generealization gap be made smaller.

Given these constraints, this experiment is conducted only to indicate that this aspect of model performance also needs to be looked into more deeply in the future. This experiment though purely indicative in nature, it is a step to show that building generalized models for certain crystal families or all types of crystal structures is very important, because novel material development will mostly always contain samples that have never been seen by an ML model. As a specific example, if in the future there needs to be analysis done on a rare class of perovskites (pseudohalid/mixed halide perovskites), and there is not enough data to train a machine learning model, then a pre-existing model that is known to generalize better might be more useful in studying these compounds.

It also has to taken into account that more perovskite semiconductors and insulators need to be added to our dataset, either by employing DFT or experiments. A larger dataset will irrevocably assist in building a robust ML model, one that is more accurate and can understand the underlying data distribution and hence generalize on novel materials better.

5. Conclusion

We have collated a dataset that encompasses a wide variety of perovskite compounds ranging from organometallic halid perovskites to 2D perovskites and their corresponding bandgaps. The structures of the perovskites can be used to calculate any important property of a perovskite using Density Functional Theory or ab initio molecular dynamics. The corresponding bandgaps have been collated for training a graph neural network model, MEGNet, to accurately predict the property. Certain studies were made on the model to understand the nature of the data and also interpret the model. Furthermore, the model was inserted into a pipeline, that allows any researcher to try various combinations of inorganic perovskites to obtain its bandgap. Both the pipeline and the dataset have been made available on our github repository.

We underscore the importance of adding onto our dataset more types of perovskites, specifically double perovskites, mixed halide perovskites and also to ensure there are more semiconductor perovskites. Such a robust dataset is key to build a model that performs well. We also highlight the need for building models that will generalize better for novel materials.

As future work, the pipeline needs to be altered for including hybrid organic–inorganicperovskites, and ML models that predict lattice constants for perovskites can be used to replace the DFT calculators, as this will speed up the process to predict bandgaps. Though we have chosen MEGNet to be our model, any future graph neural network model that performs better can replace our predictor.

This work is a step in the direction of highlighting the importance of building models that generalize well, creating easy workflows for using researchers new to computational material science to easily shortlist and characterize compounds, and finally the need for robust datasets of considerable size to get more accurate models that perform better for novel materials.

CRediT authorship contribution statement

Pravan Omprakash: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Project administration, Supervision, Investigation. **Bharadwaj Manikandan:** Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing, Project administration, Supervision. **Ankit Sandeep:** Methodology, Software, Formal analysis, Data curation, Writing - original draft, Writing - review & editing. **Romit Shrivastava:** Methodology, Software, Formal analysis, Data curation, Writing - original draft, Investigation. **Viswesh P.:** Methodology, Software, Formal analysis, Data curation, Writing - original draft, Investigation. **Devadas Bhat Panemangalore:** Writing - review & editing, Project administration, Supervision, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M.A. Green, A. Ho-Baillie, H.J. Snaith, The emergence of perovskite solar cells, Nature Photonics 8 (2014) 506–514.

[2] M.V. Kovalenko, L. Protesescu, M.I. Bodnarchuk, Properties and potential optoelectronic applications of lead halide perovskite nanocrystals, Science 358 (2017) 745–750.

[3] T. Tiedje, E. Yablonovitch, G.D. Cody, B.G. Brooks, Limiting efficiency of silicon solar cells, IEEE Transactions on Electron Devices 31 (1984) 711–716.

[4] F. Sahli, J. Werner, B.A. Kamino, M. Bräuninger, R. Monnard, B. Paviet-Salomon, L. Barraud, L. Ding, J.J.D. Leon, D. Sacchetto, et al., Fully textured monolithic perovskite/silicon tandem solar cells with 25.2% power conversion efficiency, Nature Materials 17 (2018) 820–826.

[5] H.-S. Kim, J.-Y. Seo, N.-G. Park, Material and device stability in perovskite solar cells, ChemSusChem 9 (2016) 2528–2540.

[6] W. Kohn, A.D. Becke, R.G. Parr, Density functional theory of electronic structure, The Journal of Physical Chemistry 100 (1996) 12974–12980.

[7] K. Burke, Perspective on density functional theory, The Journal of Chemical Physics 136 (2012), 150901.

[8] J. Wei, X. Chu, X.-Y. Sun, K. Xu, H.-X. Deng, J. Chen, Z. Wei, M. Lei, Machine learning in materials science, InfoMat 1 (2019) 338–358.

[9] G. Pilania, A. Mannodi-Kanakkithodi, B. Uberuaga, R. Ramprasad, J. Gubernatis, T. Lookman, Machine learning bandgaps of double perovskites, Scientific Reports 6 (2016) 1–10.

[10] G. Pilania, P.V. Balachandran, C. Kim, T. Lookman, Finding new perovskite halides via machine learning, Frontiers in Materials 3 (2016) 19.

[11] P.V. Balachandran, B. Kowalski, A. Sehirlioglu, T. Lookman, Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning, Nature Communications 9 (2018) 1–9.

[12] T. Xie, J.C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Physical Review Letters 120 (2018), 145301.

[13] J. Im, S. Lee, T.-W. Ko, H. Kim, Y. Hyon, H. Chang, Identifying pb-free perovskites for solar cells by machine learning, npj Computational Materials 5 (2019) 8.

[14] Z. Li, Q. Xu, Q. Sun, Z. Hou, W.-J. Yin, Thermodynamic stability landscape of halide double perovskites via high-throughput computing and machine learning, Advanced Functional Materials 29 (2019) 1807280.

[15] G. Fan, K. Han, G. He, Time-dependent density functional-based tight-bind method efficiently implemented with openmp parallel and gpu acceleration, Chinese Journal of Chemical Physics 26 (2014) 635.

[16] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning, Nature Communications 9 (2018) 1–8.

[17] S. Sun, N.T. Hartono, Z.D. Ren, F. Oviedo, A.M. Buscemi, M. Layurova, D.X. Chen, T. Ogunfunmi, J. Thapa, S. Ramasamy, et al., Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis, Joule 3 (2019) 1437–1451.

[18] S.-Y. Louis, Y. Zhao, A. Nasiri, X. Wang, Y. Song, F. Liu, J. Hu, Graph convolutional neural networks with global attention for improved materials property prediction, Physical Chemistry Chemical Physics 22 (2020) 18141–18148.

[19] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. Ong, Graph networks as a universal machine learning framework for molecules and crystals, Chemistry of Materials 31 (2019) 9.

[20] P.W. Battaglia, J.B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, et al., Relational inductive biases, deep learning, and graph networks, arXiv preprint arXiv: 1806.01261 (2018) 40.

[21] S.Y. Louis, Y. Zhao, A. Nasiri, X. Wong, Y. Song, F. Liu, J. Hu, Global attention based graph convolutional neural networks for improved materials property prediction, arXiv preprint arXiv:2003.13379 (2020) 11.

[22] I.E. Castelli, J.M. García-Lastra, K.S. Thygesen, K.W. Jacobsen, Bandgap calculations and trends of organometal halide perovskites, APL Materials 2 (2014), 081514.

[23] C. Kim, T. Huan, S. Krishnan, R. Ramprasad, A hybrid organic-inorganic perovskite dataset, Scientific Data 4 (2017) 11.

[24] I.E. Castelli, D.D. Landis, K.S. Thygesen, S. Dahl, I. Chorkendorff, T.F. Jaramillo, K. W. Jacobsen, New cubic perovskites for one- and two-photon water splitting using the computational materials repository, Energy & Environmental Science 5 (2012) 9034–9043.

[25] K. Kuhar, A. Crovetto, M. Pandey, K.S. Thygesen, B. Seger, P.C.K. Vesborg, O. Hansen, I. Chorkendorff, K.W. Jacobsen, Sulfide perovskites for solar energy conversion applications: computational screening and synthesis of the selected compound lays3, Energy & Environmental Science 10 (2017) 2579–2593.

[26] I.E. Castelli, K.S. Thygesen, K.W. Jacobsen, Calculated optical absorption of different perovskite phases, Journal of Materials Chemistry A 3 (2015) 12343–12349.

[27] I.E. Castelli, J.M. García-Lastra, F. Hüser, K.S. Thygesen, K.W. Jacobsen, Stability and bandgaps of layered perovskites for one- and two-photon water splitting, New Journal of Physics 15 (2013), 105026.

[28] Y. Zhuo, A. Mansouri Tehrani, J. Brgoch, Predicting the band gaps of inorganic solids by machine learning, The Journal of Physical Chemistry Letters 9 (2018) 1668–1673. PMID:29532658.

[29] J.C. Stanley, F. Mayr, A. Gagliardi, Machine learning stability and bandgaps of lead-free perovskites for photovoltaics, Advanced Theory and Simulations 3 (2020) 1900178.

[30] V. Gladkikh, D.Y. Kim, A. Hajibabaei, A. Jana, C.W. Myung, K.S. Kim, Machine learning for predicting the band gaps of abx3 perovskites from elemental properties, The Journal of Physical Chemistry C 124 (2020) 8905–8918.

[31] K. Takahashi, L. Takahashi, I. Miyazato, Y. Tanaka, Searching for hidden perovskite materials for photovoltaic systems by combining data science and first principle calculations, ACS Photonics 5 (2018) 771–775.

[32] W.A. Saidi, W. Shadid, I.E. Castelli, Machine-learning structural and electronic properties of metal halide perovskites using a hierarchical convolutional neural network, npj Computational Materials 6 (2020) 1–7.

[33] Articial neural networks for accurate prediction and analysis of perovskite bandgaps, ECS Meeting Abstracts (2019).

[34] I.E. Castelli, M. Pandey, K.S. Thygesen, K.W. Jacobsen, Band-gap engineering of functional perovskites through quantum confinement and tunneling, Physical Review B 91 (2015), 165309.

[35] C. Kim, G. Pilania, R. Ramprasad, Machine learning assisted predictions of intrinsic dielectric breakdown strength of abx3 perovskites, The Journal of Physical Chemistry C 120 (2016) 23.

[36] D.D. Landis, J.S. Hummelshoj, S. Nestorov, J. Greeley, M. Dulak, T. Bligaard, J. K. Norskov, K.W. Jacobsen, The computational materials repository, Computing in Science & Engineering 14 (2012) 51–57.

[37] S.P. Ong, W.D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K.A. Persson, G. Ceder, Python materials genomics (pymatgen): A robust, open-source python library for materials analysis, Computational Materials Science 68 (2013) 314–319.

[38] W. Kaminsky, T. Snyder, J. Stone-Sundberg, P. Moeck, One-click preparation of 3d print files (*. stl,*. wrl) from*. cif (crystallographic information framework) data using cif2vrml, Powder Diffraction 29 (2014) S42–S47.

[39] Y. Chen, Y. Sun, J. Peng, J. Tang, K. Zheng, Z. Liang, 2d ruddlesden–popper perovskites for optoelectronics, Advanced Materials 30 (2018) 1703487.

[40] J. Deng, J. Li, Z. Yang, M. Wang, All-inorganic lead halide perovskites: a promising choice for photovoltaics and detectors, Journal of Materials Chemistry C 7 (2019) 12415–12440.

[41] Progress toward stable lead halide perovskite solar cells, Joule 2 (2018) 1961–1990.

[42] W. Ke, M.G. Kanatzidis, Prospects for low-toxicity lead-free perovskite solar cells, Nature Communications 10 (2019) 965.

[43] C. Chen, W. Ye, Y. Zuo, C. Zheng, S.P. Ong, Supplementary information graph networks as a universal machine learning framework for molecules and crystals (2018).

[44] A. Dunn, Q. Wang, A. Ganose, D. Dopp, A. Jain, Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm, npj Computational Materials 6 (2020).

[45] L. Torgo, R.P. Ribeiro, B. Pfahringer, P. Branco, Smote for regression, in: Portuguese Conference on Artificial Intelligence, Springer, 2013, pp. 378–389.

[46] C. Wei, J.D. Lee, Q. Liu, T. Ma, Regularization matters: Generalization and optimization of neural nets vs their induced kernel, Advances in Neural Information Processing Systems (2019) 9712–9724.

[47] G. Kieslich, S. Sun, A.K. Cheetham, An extended tolerance factor approach for organic–inorganic perovskites, Chemical Science 6 (2015) 3430–3433.

[48] Y. Fu, M.P. Hautzinger, Z. Luo, F. Wang, D. Pan, M.M. Aristov, I.A. Guzei, A. Pan, X. Zhu, S. Jin, Incorporating large a cations into lead iodide perovskite cages: Relaxed goldschmidt tolerance factor and impact on exciton–phonon interaction, ACS Central Science 5 (2019) 1377–1386.

[49] C.J. Bartel, C. Sutton, B.R. Goldsmith, R. Ouyang, C.B. Musgrave, L.M. Ghiringhelli, M. Scheffler, New tolerance factor to predict the stability of perovskite oxides and halides, Science advances 5 (2019) eaav0693.

[50] T. Sato, S. Takagi, S. Deledda, B.C. Hauback, S.-I. Orimo, Extending the applicability of the goldschmidt tolerance factor to arbitrary ionic compounds, Scientific Reports 6 (2016) 23592.

[51] A.E. Fedorovskiy, N.A. Drigo, M.K. Nazeeruddin, The role of goldschmidt's tolerance factor in the formation of a2bx6 double halide perovskites and its optimal range, Small Methods 4 (2020) 1900426.

[52] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Physical Review B 59 (1999) 1758–1775.

[53] J. Enkovaara, C. Rostgaard, J.J. Mortensen, J. Chen, M. Dułak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H.A. Hansen, H.H. Kristoffersen, M. Kuisma, A.H. Larsen, L. Lehtovaara, M. Ljungberg, O. Lopez-Acevedo, P.G. Moses, J. Ojanen, T. Olsen, V. Petzold, N.A. Romero, J. Stausholm-Møller, M. Strange, G. A. Tritsaris, M. Vanin, M. Walter, B. Hammer, H. Häkkinen, G.K.H. Madsen, R. M. Nieminen, J.K. Nørskov, M. Puska, T.T. Rantala, J. Schiøtz, K.S. Thygesen, K. W. Jacobsen, Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method, Journal of Physics: Condensed Matter 22 (2010), 253202.

[54] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Physical Review Letters 77 (1996) 3865–3868.

[55] H.J. Monkhorst, J.D. Pack, Special points for brillouin-zone integrations, Physical Review Letters 13 (1976) 5188–5192.

[56] A. Majid, A. Khan, G. Javed, A.M. Mirza, Lattice constant prediction of cubic and monoclinic perovskites using neural networks and support vector regression, Computational Materials Science 50 (2010) 363–372.

[57] S.P. Ong, W.D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K.A. Persson, G. Ceder, Python materials genomics (pymatgen): A robust, open-source python library for materials analysis, Computational Materials Science 68 (2013) 314–319.