Predicting Perovskite Bandgap and Solar Cell Performance with Machine Learning

Elif Ceren Gok[1], Murat Onur Yildirim[1], M.P.U. Haris[2], Esin Eren[3,4], Meenakshi Pegu[2], Naveen Hirandu[2], Peng Huang[2], Samrana Kazim [2,5], Aysegul Uygun Oksuz[4], Shahzada Ahmad [2,5*]

[1]Department of Mathematics and Computer Science, Engineering Faculty, Eindhoven University of Technology, 5612 AZ, Eindhoven, The Netherlands

[2] BCMaterials, Basque Center for Materials, Applications, and Nanostructures, UPV/EHU Science Park, 48940, Leioa, Spain

[3] Department of Energy Technologies, Innovative Technologies Application and Research Center, Suleyman Demirel University, 32260, Isparta, Turkey.

[4] Department of Chemistry, Faculty of Arts and Science, Suleyman Demirel University, 32260 Isparta, Turkey

[5] IKERBASQUE, Basque Foundation for Science, Bilbao, 48013, Spain

Abstract

Perovskites as a semiconductor are of profound interest and arguably, the investigation on the distinctive perovskite composition is paramount to fabricate efficient devices and solar cells. We probed the role of anion and cations and their impact on optoelectronic and photovoltaic properties. We report a machine learning approach to predict the bandgap and power conversion efficiency by employing eight different perovskites compositions. The predicted solar cell parameters validate the experimental data. The adopted Random forest model presented a good match with high R^2 scores of >0.99 and >0.82 for predicted absorption and *J-V* data sets respectively and showed minimal error rates with precise prediction of bandgap and power conversion efficiencies. Our results suggest that the machine learning technique is an innovative approach to aid the preparation of perovskite and can accelerate the commercial aspects of perovskite solar cells without fabricating working devices and minimizes the fabrication steps and save cost.

E-mail addresses of Corresponding authors: m.o.yildirim@tue.nl; erenes82@gmail.com; shahzada.ahmad@bcmaterials.net

1. Introduction

Perovskites as semiconductors have gained enormous interest in diverse scientific fields, due to their ease of processing, tuneability, and unparalleled semiconducting features[1]. Hybrid perovskite is widely being used in a variety of applications including solar cells, light-emitting diodes, lasers, and photodetectors due to its longer electron and hole diffusion lengths and higher carrier mobility, and wide tunable bandgap (E_g).[1] The perovskite structure is represented by ABX_3[2,3], where A is an organic cation, B is metal, and X is a halogen anion (Figure 1a). The rapid astonishing advances in the power conversion efficiency (PCE) of perovskite solar cells (PSCs) led to a rise in the PCE from 3.8 to 25.5% and currently occupies the center stage for photovoltaic research and development.[4,5] In the current decade, the perovskites-based semiconductors have witnessed a surge of increment in terms of scientific publications and are being intensively investigated.[1].

Routinely, photo-active perovskites are developed through a trial-and-error method, where the target properties are achieved by continual synthesis and characterizations. To accelerate the materials discovery path and overcome the laborious and expensive effort, computational approaches like Monte Carlo simulation,[6] molecular dynamics[7], and Density functional theory[8] (DFT) are being used to compute the materials that can deliver promising optoelectronic properties. Nevertheless, such simulation methods are designed for specific systems and their employment in a wide variety of materials makes the task demanding along with its high cost and tedious process. In this context, machine learning (ML) has played a substantial role to discover high-performance materials.[9,10] ML is a data-driven approach that combines with experimental datasets to predict concealed information and trends[2]. ML approaches established as influential for properties and performance prediction of materials, which in turn can expedite the material exploration by minimizing the tasks for proposing potentially promising structures.[11] The accuracy of the ML approach is comparable with the DFT calculations and arguably viable to design the materials even from small data sets, while the small and uneven distributions are insufficient for DFT calculations.[12,13] The bandgap tuning and prediction are of significance for perovskite applications in light emission and harvesting. The bandgap of the perovskites can be varied from 1.5 − 3.2 eV with the anions selection, varying composition, and the A-site cations (Cs, formamidinium (FA), methylammonium (MA)[3], etc. A random forest (RF) model has been used to predict the bandgap of Li and Na-based perovskite using 18 physical descriptors, and 9328 types of materials with ideal bandgaps to capture solar light were estimated.[14] Similarly, a linear regression model was developed to predict the bandgap of the mixed halide hybrid perovskite with higher accuracy (RMSE of 0.05 eV)^3. Zheng and co-workers compared different ML models, ie, RF, Ridge regression (RR), and support vector regression (SVR), to predict four target variables including perovskite bandgap from seven descriptors, and the authors noted the high accuracy of the RF model.[15] Compositional engineering of the perovskite is an effective approach to fabricate efficient PSCs. It is imperative to track how the cationic and anionic engineering of perovskite will influence the optical bandgap and impact device performance. Unraveling such, information will be crucial for understanding and is paramount to predict or design materials with added merits. To our knowledge, most of the ML approaches for PSCs have been carried out using the literature data as the input variables. Arguably, the data from different laboratory conditions will increase the error factor and may end up with less accuracy in the prediction. Here we use the descriptor data sets obtained from a single (our) laboratory conditions which could improve the performance of the ML model.

Here, we applied the ML approach in two different steps: (i) to predict the bandgap and (ii) PSCs performance by employing eight different perovskites. Firstly, we derived the bandgap of perovskites from Tauc plots (UV-Vis spectroscopy) using both the experimental and ML approaches. Secondly, we built the model for *J-V* spectra prediction to evaluate the PSCs performance. Our work suggests solar cells performance prediction and eliminates the need to fabricate working devices, which in turn save costs and avoid environmental hazards.

2. Experimental

Materials: All chemicals were purchased from Sigma Aldrich unless and otherwise stated and were used as received without any further purification. CsI, MA, FA, PbI_2 was procured from TCI, while Chlorobenzene (CB), isopropanol (IPA, 99.9 %), anhydrous dimethyl sulfoxide (DMSO, 99.8 %), and *N, N*-dimethylformamide (DMF. 99.8%) were purchased from Acros Organics. Perovskite precursors were purchased from Dyesol, while PbI_2 and CsI_2 were procured from Tokyo Chemical Industry (TCI). [60]PCBM >99.5 % and Bathocuproine (BCP) were purchased from Solenne BV and TCI respectively.

Perovskites: Eight different types of perovskites layers were deposited as follows

RbCsFAMAPI: The quadruple-cation perovskite precursor solution was prepared using FAI (1 M), PbI_2 (1.1 M), MABr (0.2 M), and $PbBr_2$ (0.22 M) in 1 mL of anhydrous solvents mixture of DMF and DMSO with 4:1 (v/v) ratio. 42 μL of 1.5 M CsI solution and 42 μL of 1.5 M RbI solutions in DMSO and DMSO: DMF (1:4 v/v) solvents were added to the above solution and stirred for 1 h. The precursor solution was spin-coated in a two-step spin-coating program (1000 rpm and 6000 rpm for 10 and 30 s, respectively). 112 µL of chlorobenzene was dripped at 10 s before ending the second spin step followed by annealing at 100 °C for 1 h.

CsFAMAPI: The triple-cation perovskite [$Cs_{0.1}(FAPbI_3)_{0.81}(MAPbBr_3)_{0.09}$] precursor solution was prepared containing CsI (0.10M), FAI (1.05 M), PbI_2 (1.24 M), MABr (0.12 M) and $PbBr_2$ (0.12 M) in 1mL of anhydrous solvents mixture of *N,N*-dimethylformamide (DMF) and dimethylsulfoxide (DMSO) with 4:1 (v/v) ratio. The precursor solution was spin-coated in a two-step spin-coating program (1000 rpm and 6000 rpm for 10 and 30 s, respectively). 112 µL of chlorobenzene was dripped at 10 s before ending the second spin step followed by annealing at 100 °C for 1 h.

FAPI+MAPBr: Precursor solution was prepared by mixing FAI (1 M), PbI_2 (1.2 M), MABr (0.2 M), and $PbBr_2$ (0.2 M) in 1mL of anhydrous solvents mixture of *N, N*-dimethylformamide (DMF), and dimethylsulfoxide (DMSO) with 4:1 (v/v) ratio. The precursor solution was spin-coated in a two-step spin-coating program (1000 rpm and 6000 rpm for 10 and 30 s, respectively). 112 µL of chlorobenzene was dripped at 10 s before ending the second spin step followed by annealing at 100 °C for 1 h.

MAPI: The $MAPbI_3$ precursor solution was realized by dissolving an equimolecular amount of MAI and PbI_2 (1.2M) in DMSO solvent. The precursor solution was spin-coated in a two-step spin-coating program (1000 rpm and 4000 rpm for 10 and 30 s, respectively). 112 µL of chlorobenzene was dripped at 10 s before ending the second spin step followed by annealing at 100 °C for 1 h.

FAPI and CsFAPI: Instead of conventional precursor materials, pre-synthesized non-perovskite yellow powders were employed as the precursor materials for FAPI and CsFAPI precursor solution, and the powder precursor synthesis was reported in our previous work[4].1.25 M precursor solutions were prepared by dissolving 791.25 mg of δ-$FAPbI_3$ and 800 mg of δ-$CsFAPbI_3$ powders in a 1 mL anhydrous solvent mixture of DMF and DMSO with a 4:1 (v/v) ratio. The FAPI and CsFAPI perovskites were fabricated by spin coating the precursor solutions at 1000 for 5 s and 5000 rpm for 20 s. 100 µL chlorobenzene was dripped at the final 5 s of spinning and the FAPI and CsFAPI thin films were annealed at 150 °C for and 80 °C respectively.

MAPI-Cl: The $MAPbI_{3-x}Cl_x$ perovskite films were fabricated by a two-step deposition method[16]. The PbI_2 solution was prepared by the dissolution of PbI_2 in DMF and stirred at 70 °C for 12 h. The mixed cation solution was prepared by dissolving MAI and MACl with the concentrations of 50 mg and 5 mg/mL in 2-propanol (IPA), respectively. The PbI_2 films were spun at 4,500 rpm for 20 s using warm PbI_2 solution at 70 °C, and then a drop of mixed cation solution was dropped on the center of spin-coated film for 30 s. The as-prepared samples were annealed at 100 °C for 3 min.

FAMAPI-Br: The $FAMAPbI_{3-x}Br_x$ films were also fabricated by a two-step deposition method. Firstly, 1.3 M PbI_2 dissolved in a mixed solvent (DMF/DMSO = 9.5/0.5) was dissolved at 70 °C for overnight. The warm PbI_2 was spin-coated at 4000 rpm for 20 s, then a drop of the mixed organic solution with FAI/ MABr/ MACl =60 mg/ 6 mg/ 6 mg) in 1 mL isopropanol was added on the spinning substrate for 30 s. The as-prepared samples were annealed at 150 °C for 15 min under ambient conditions with 30-40% RH.

Device fabrications: Both types of architects were adopted for device fabrication *n-i-p* type: The solar cells were fabricated on commercial laser-etched FTO glass electrodes (10 Ω/sq, NSG). All of the electrodes were cleaned by sonication in sequence with Hellmanex II solution, Milli-Q water, acetone, and 2-propanol for 20 min each (pre-cleaning). The cleaned substrates were dried with a stream of compressed air and were further treated by UV-ozone for 15 min before device fabrication. For the *n-i-p* (1) PSC, The SnO_2 electron transporting layer is prepared by spin-coating the 3% wt% SnO_2 nanoparticles (Alfa Aesar) at 5, 000 rpm for 30 s, and then post-heated at 150 °C for 15 min. For the *n-i-p* (2), compact TiO_2 (c-TiO_2) layer was deposited using spray pyrolysis at 500 °C employing 1 mL of titanium diisopropoxide bis(acetylacetonate) precursor solution (75% in IPA) in 19 mL of pure ethanol using oxygen as the carrier gas, followed by annealing for another 30 min at 500 °C to acquire the anatase phase. SnO_2 quantum dots (SnO_2-QD) synthesized by a previously reported method[17]was spun coated on FTO:c-TiO_2 substrate followed by annealing at 150 °C for 45 minutes. On the other hand, for the *n-i-p* (3), the TiO_2 mesoporous (mp-TiO_2) layer (1:8 w/v in ethanol) was spin-coated over FTO:c-TiO_2 substrate at 4000 rpm with 2000 rpm/s acceleration for 30 s, followed by progressive heating steps until 500 °C for 30 min. Then the substrates were treated with UV-ozone for 30 min and transferred immediately to the argon-filled glove box at the room and the different perovskite layers were deposited as discussed above.

For the HTM layer, 70- and 60-mM Spiro-OMeTAD was prepared by dissolving the desired amount of material in 1mL chlorobenzene. Doping was achieved by the addition of 4-tert-butylpyridine (38.4 and 28.8 μL for 70 and 60 mM, respectively) and bis(trifluoromethylsulfonyl)imide lithium salt solution with a concentration of 520 mg/mL (21.1 and 17.5 μL for 70 and 60 mM respectively). 60 mM Spiro-OMeTAD solution was used as HTM for FAPI and CsFAPI perovskites and 70 mM solution was employed for all other perovskites. 35μL of HTM solution was dropped on the perovskite layer and were spin-coated at 4000 rpm for 20 s. Au electrode (80 nm) was thermally evaporated under a pressure of 2 × 10^{-6} Pa to complete the device fabrication.

p-i-n type: The PEDOT-PSS film was spin-coated on the pre-cleaned and UV-Ozone treated ITO substrates at 5,000 rpm for 30 s in air, and post-heating at 150 °C for 15 min. ITO/PEDOT-PSS substrates were transferred to the glove box for the perovskite layer deposition. The electron transporting layer (10 mg/mL $PC_{61}BM$ in chloroform) was spin-coated on the perovskite layer at 1,200 rpm for 30 s, then the thin BCP film was spin-coated on the samples at 5,000 rpm for 30 s using 0.5 mg/mL isopropanol. Finally, the 100 nm thick Ag electrode was deposited by thermal evaporation.

Prediction Models

ML can unravel conceal patterns and generate representative models from the data without assigning specific instructions to the machines.[18] ML focuses on prediction by employing general-purpose learning algorithms to uncover patterns in occasionally complex and cumbersome data sets. Even when data is collected without a tightly controlled experimental design and in the context of complex nonlinear interactions, it returns effective results. On the other hand, statistical approaches emphasis on inferences that are performed through the design and fitting of a project-specific probability model.[19] In the chemical processes, typically each experiment builds data to explore and the task of revealing the patterns become demanding when the number of experiments and data set increases. Arguably, ML develop as an efficient means to monitor the data in the chemical processes. Numerous ML-based approaches can be used for processing different data sets from simple to complex. Random Forest consists of several individual tree structures is a collective method that could be run for different tasks such as classification and regression. When the workload is to classify, the random forest interprets each individual tree and makes decisions based on majority voting while the mean or average of the individual trees is returned for the regression task. The basic principle behind random forest is the wisdom of crowds which is a simple yet effective approach. Any of the individual constituent models would be outperformed by a large number of reasonably uncorrelated trees working as a committee. The key point is to have low correlation between the trees in order to have a more generalizable model. Therefore, the RF is an attractive model that validates its performance with a flexible structure. In this method, we constructed randomly more than one decision tree (Figure 1b), to allow robust prediction and the default hyperparameter configuration employed is already defined by sklearn library. For instance, number of trees, min_samples_split and min_samples_leaf was selected as 100, 2, and 1 accordingly while fixing the seed of random_state to have reproducible model. Combining multiple randomly structured decision trees into one model results in enhanced predictions.[20]

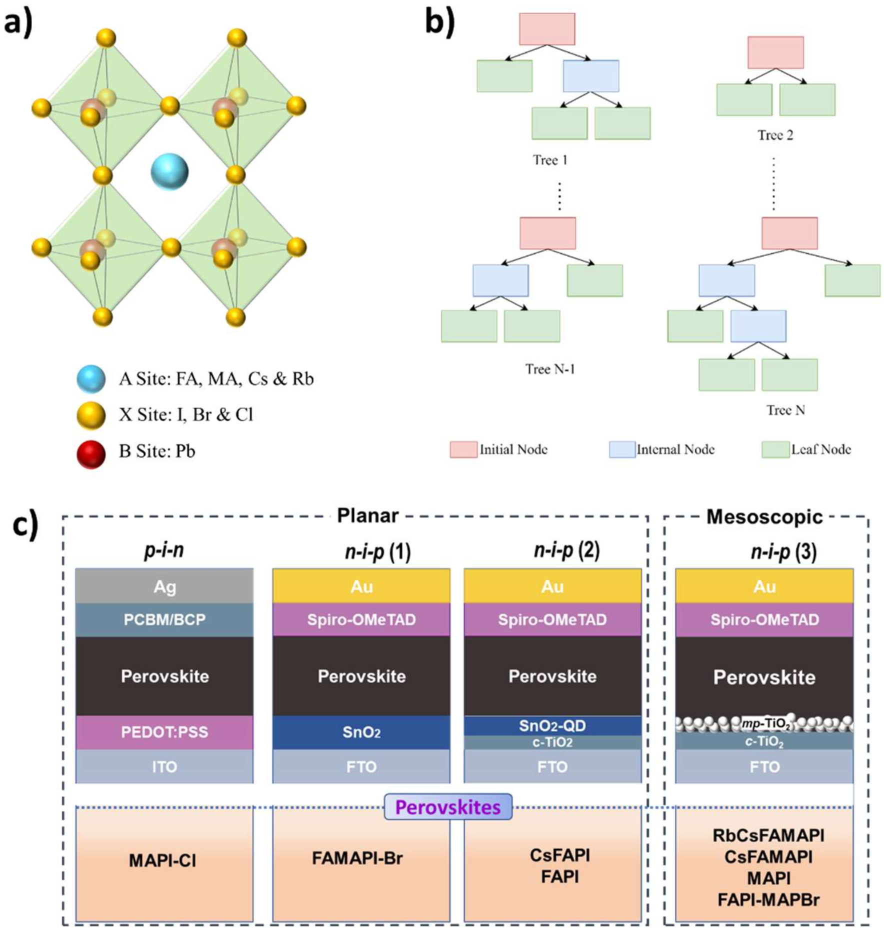

Figure 1. a) Crystal structure of hybrid organic-inorganic perovskite, b) illustration of the Random Forest Model, and c) schematic representation of PSCs employed in the study.

To accurately evaluate model performance we adopted both k-fold cross-validation and the typical train test split. The presented datasets (Table 1) were firstly split into 80% train set and 20% test set and k-fold cross-validation were applied on the train set to note the overfitting. The final evaluation measure was obtained on a 20% test set. For this, we split the data into k = 5 pieces and, at each iteration, k-1 number of sets are employed to train the model while the remaining piece is utilized for the evaluation.[21].

Table 1. Train/test size for UV and *J-V* datasets

Material	UV (# of samples)		*J-V* (# of samples)	
	Train Set	Test Set	Train Set	Test Set
FAPI	180	46	56	15
CsFAPI	183	46	74	19
RbCsFAMAPI	183	16	76	20
CsFAMAPI	187	47	76	19
FAPIMABr	183	46	76	19
MAPI	186	47	76	19
MAPICI	152	39	40	10
FAMAPIBr	183	46	77	20

To measure the success of the regression models certain performance metrics are required such as R^2 and root mean squared error (RMSE). When computing the prediction error, RMSE assigns equal weight to each data point, whereas R^2 is more sensitive to outliers. R^2 score in equation 1 was chosen as a performance indicator in this study to also consider outliers in the data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_{pred})^2}{\sum_{i=1}^{n}(y_i - y_{mean})^2} \qquad \text{Equation 1}$$

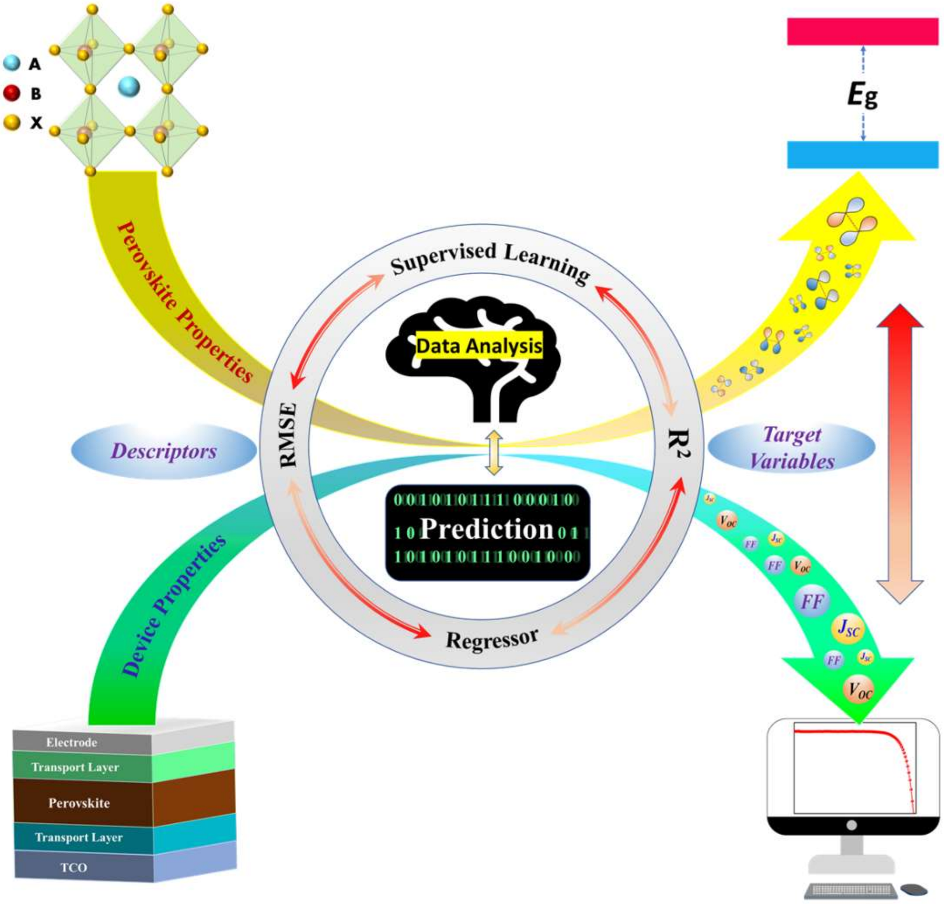

Figure 2: The schematic diagram and inputs, outputs, and processing parameters.

2. Results and Discussions

The functional relationship between descriptors (UV-Vis absorption, *J-V,* and EQE curves) and the target variables (optical bandgap and photovoltaic parameters) has been decoded with the RF regression method. To probe the influence of A and X site variations in the lead halide perovskites, we selected eight different perovskites (Figure 1a) including the typical and most studied $MAPbI_3$, mixed perovskites, and the $FAPbI_3$ for analysis. Hereafter we have named the different perovskite structures as RbCsFAMAPI [$Rb_{0.5}Cs_{0.5}(FAPbI_3)_{0.83}(MAPbBr_3)_{0.17}$], CsFAMAPI [$Cs_{0.1}(FAPbI3)_{0.81}(MAPbBr_3)_{0.09}$], CsFAPI [$Cs_{0.1}FA_{0.9}PbI_3$], FAPI [$FAPbI_3$], MAPI [$MAPbI_3$], MAPI-Cl [$MAPbI_{3-x}Cl_x$], FAPI+MAPBr [$(MAPbBr_3)_{0.15}(FAPbI_3)_{0.85}$] and FAMAPI-Br [$FAMAPbI_{3-x}Br_x$].

Optical properties prediction

We evaluated the influence of stoichiometric alterations on the optical properties through absorption spectroscopy. The experimental and predicted UV-Vis absorption spectra of eight different perovskites are shown (Figure 3). From the experimental data, the absorption onsets of RbCsFAMAPI, CsFAMAPI, CsFAPI, FAPI, MAPI, MAPI-Cl, FAPI+MAPBr, and FAMAPI-Br were calculated as ~777 nm, ~790 nm, ~812 nm, ~848 nm, ~ 804 nm, ~784 nm, ~778 nm, ~835 nm, respectively. The absorption onset of perovskites has a strong negative correlation with the electronegativities of the halide components, i.e., the higher electronegative component shows lower absorption onset. The blue-shifted absorption onset of MAPI-Cl compared to the MAPI suggests the Cl inclusion. On the other hand, the A site substitutions were evaluated with their lattice constants and represent a positive correlation with the absorption onset. [22]. The extended absorption onset of FAPI than MAPI has been attributed to the higher lattice constant of FA over MA and expectedly, the other mixed perovskites showed the absorption onset within the range. However, in the case of quadruple-cation, with the addition of Rb cation (higher lattice constant), the absorption onset of the CsFAMAPI layers is shifted to a lower wavelength (~13 nm), suggesting a higher bandgap. We ascribed this blue shift related to the surface of perovskite due to band filling and/or reduced surface traps.[23] The performance of the RF model is indicated with a high R^2 score (Table 2), all the studied perovskites showed an exceptional R^2 value of >0.99 indicating the strong correlation between predicted and experimental curves. After evaluating the R^2 scores of each material, to demonstrate the model's generalization performance, the average R^2 score and standard deviation based on the eight perovskites were calculated as 0.9979 and 0.0021 respectively. The high average R^2 score and low standard deviation suggest the remarkable efficacy of the RF model over various material types. Further, the Tauc plots were used on experimental and predicted datasets to estimate the optical bandgap of perovskites, by plotting the $(\alpha h\nu)^2$ vs $h\nu$[24], (Figure S1-S8) and the resultant E_g values are tabulated (Table 2).

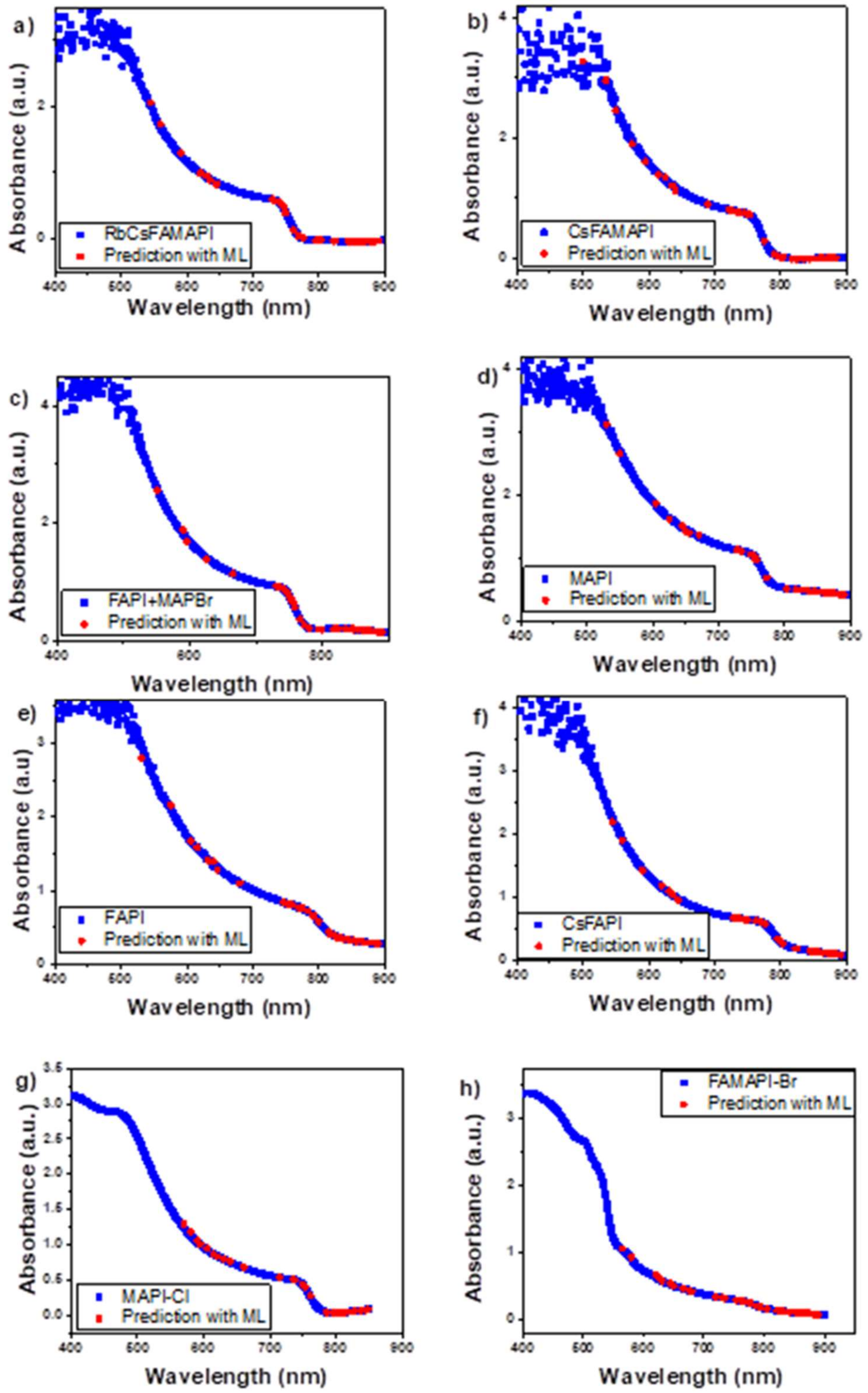

Figure 3. Measured (blue) and RF model simulated (red) UV-Vis spectra of perovskites layers for (a) RbCsFAMAPI, b) CsFAMAPI, c) FAPI+MAPBr, d) MAPI, e) FAPI, f) CsFAPI g) MAPI-Cl, and h) FAMAPI-Br.

The predicted bandgap derived from the UV, and the RF model, were consistent and displayed a low deviation of <1.4 % from the experimental results. In our case, MAPI-Cl was predicted with the higher success (0.00062 deviations) while FAMAPI-Br prediction yields the comparatively least success with a 0.01321 error rate from all the samples. Notably, MAPI yielded a top R^2 score, however, it also showed comparatively higher deviation on bandgap prediction since the RF model was obtained by the random selection of data points in the UV-Vis dataset. Bandgap calculations from the absorption data set can deviate if the randomly selected data points are not typically on the linear region. Investigating eight different perovskites with A and X sites variation, we suggest the RF model can be an appropriate model to accurately predict the optical bandgaps of lead halide perovskite. These findings signal the rational designing of the perovskite structure to push the performance.

Table 2. Performance and optical bandgap (E_g) values derived from measured and RF model-simulated data set for different perovskites.

Perovskites types	**R^2**	**Thickness (nm)**	**Predicted E_g (eV)**	**E_g (eV)**	**Deviation (%)[a]**
RbCsFAMAPI	0.9990	536.25	1.608	1.621	0.802
CsFAMAPI	0.9933	429.52	1.575	1.582	0.442
CsFAPI	0.9993	313.67	1.535	1.541	0.389
FAPI	0.9966	280.67	1.487	1.494	0.468
MAPI	0.9995	413.67	1.554	1.570	1.019
MAPI-Cl	0.9985	398	1.608	1.607	0.062
FAPI-MAPBr	0.9990	438.25	1.599	1.602	0.187
FAMAPI-Br	0.9983	322.4	1.494	1.514	1.321
AVG_R^2	0.9979				
STD_R^2	**0.0021**				

(a) Deviation=((predicted value-measured value)/measured value)*100

Photovoltaic parameters prediction

The goal of this work is to predict the performance of PSCs from the predicted current density-voltage (*J-V*) and the power-voltage characteristics. The *J-V* data from the fabricated PSCs employing the eight different perovskites as absorber layers were modeled using RF regression to test the PCE predictability. Figure 4 depicts the experimental and predicted *J-V* curves and the corresponding PCE from experimental data and the RF model are tabulated (Table 3). We employed R^2 scores to evaluate the RF efficacy for *J-V* models (Table 3), and the average R^2 score and standard deviation based on the eight perovskites were measured as 0.9010 and

0.0534, respectively. This highlight the model's generalization performance. Rational fitting of the RF regression model is achieved for all the PSCs, i.e, $0.82 < R^2 < 0.97$.

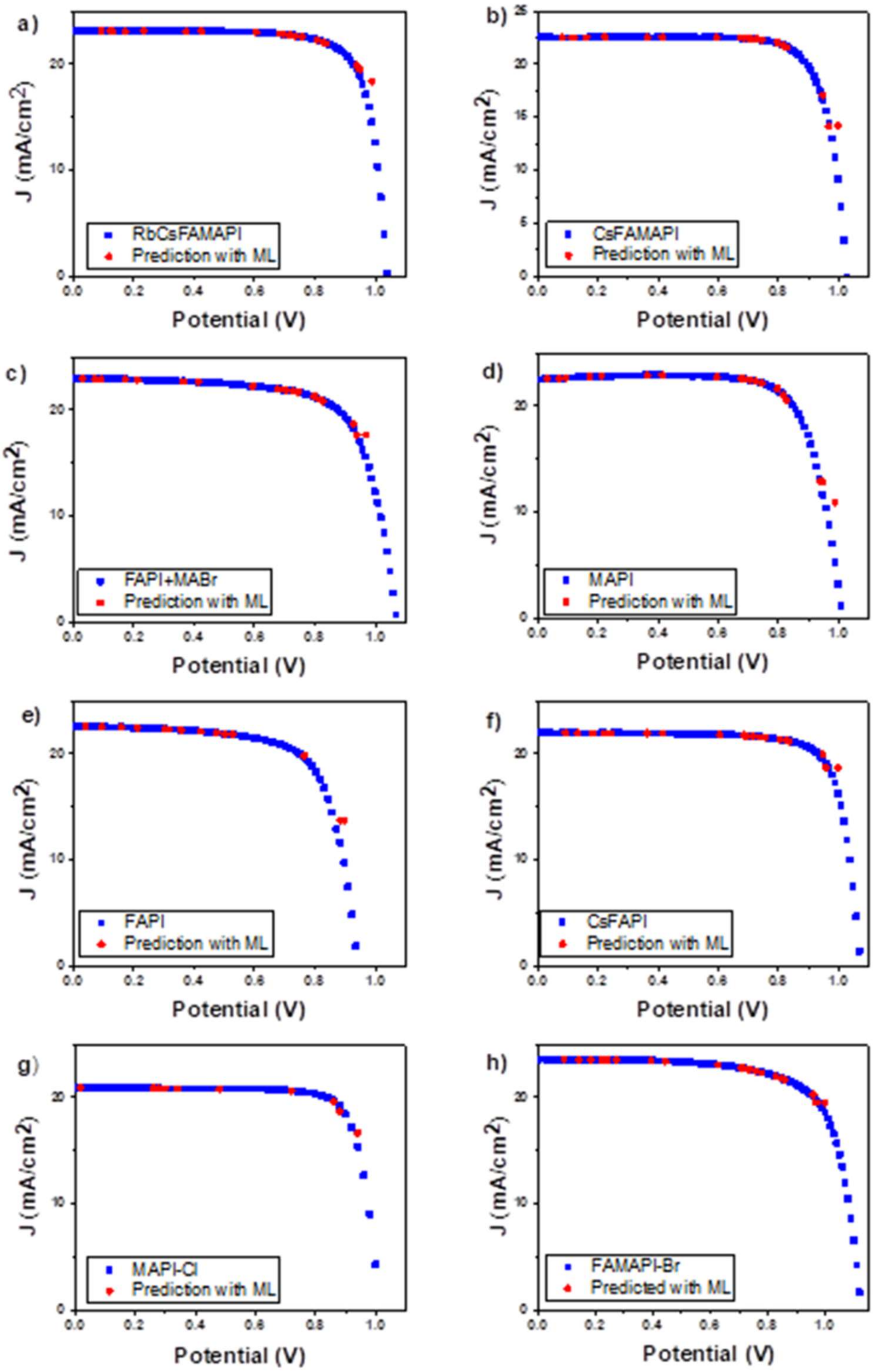

Figure 4. A comparison between the machine learning simulation and measured *J-V* graph of PSCs based on a) RbCsFAMAPI, b) CsFAMAPI, c) FAPI+MAPBr, d) MAPI, e) FAPI, f) CsFAPI, g) MAPI-Cl, and h) FAMAPI-Br.

To minimize the error factors, all the PSCs have been fabricated under the same laboratory conditions in a single laboratory, and notably, this allowed us to reach a good R^2 value. The PCEs calculated from the experimental *J-V* curves show that the FAPI based PSCs displayed the lowest PCE value of 15% and the FAMAPI-Br measured the maximum value of 19.3%. While the other fabricated PSCs fall in between, which is in agreement with the predicted PCEs by our RF regression model. It is worthy to note that MAPI and MAPI-Cl displayed maximum deviations of 0.290 and 3.176% respectively, between the measured and predicted PCEs while the other PSCs showed deviation near to ~1%. As depicted in Figure 1c, MAPI-Cl based PSC was fabricated in a *p-i-n* fashion, while the rest of the PSCs was in *n-i-p* configuration. This factor was not taken into consideration during ML. We have also not taken into account the effects of the charge transporting layers and the interfaces on the device performances to avoid complexity and this work is mainly focused on the light-harvesting layer. In comparison to the bandgap prediction, the RF regression model displayed a reduced accuracy in PCE predictions and we attribute this to the influences of charge transport layers, device architecture, interface properties, halide segregations, and induced losses. Further, we have calculated the power-voltage curves from both the experimental and RF simulated *J-V* data sets (Figure S9). The observed correlation between the experimentally calculated P-V curves with the RF simulated curve supports the commendable performance of our ML approach.

Table 3. Performance and photovoltaic parameters derived from experimental RF simulated data for different perovskite absorbers.

Perovskites types	**R^2**	**Thickness (nm)**	**Predicted PCE (%)**	**Measured PCE (%)**	**Deviation (%)[(a)]**
RbCsFAMAPI	0.8207	536.25	18.68	18.9	1.164
CsFAMAPI	0.8920	429.52	17.84	18.1	1.436
CsFAPI	0.8257	313.67	18.95	18.7	1.136
FAPI	0.9126	280.67	15.17	15	1.133
MAPI	0.9201	413.67	17.25	17.2	0.290
MAPI-Cl	0.9236	398	16.46	17	3.176
FAPI+MAPBr	0.9395	438.25	17.33	17.6	1.534
FAMAPI-Br	0.9741	322.4	19.47	19.3	0.880
AVG_R^2	**0.9010**				
STD_R^2	**0.0534**				

(a)Deviation = ((predicted value-measured value)/measured value)*100

The external quantum efficiency (EQE) measures the ratio of the number of charge carriers collected to the number of photons of given energy on light illumination. We further, assessed the performance of the proposed RF model for the EQE data set (Figure S10), it can be deduced that the overall EQE response of PSCs is in agreement with the RF simulation. This validates our findings and demonstrates the suitability of the adopted simulation model. We adopted R^2 values (Table S1) to track the performance of the RF-based EQE models and to validate the performance, the average R^2 score and standard deviation based on the eight perovskites were measured as 0.9717 and 0.0239, respectively.

Here, we have assessed the suitability of the RF regression model to predict the optical and photovoltaic properties of lead halide perovskites with A and X sites variations. The predicted and experimental PCE as a function of the predicted and calculated bandgap are plotted (Figure 5), suggesting the efficacy of our RF regression model. Though the FAPI perovskite showed the lowest bandgap of ~1.49 eV in both the experimental and RF simulation methods, it yielded the lowest PCE of 15% here, due to the method adopted for perovskite preparation. However, the FAMAPI-Br with a lower bandgap of 1.514 eV gave the highest PCE of >19%. Expectedly, CsFAPI and CsFAMAPI with a comparatively lower bandgap of <1.6 eV showed a slight decrement in PCE. In contrast the RbCsFAMAPI with a bandgap of >1.6 eV measured >18.6% PCE. Analyzing the outputs of our RF regression model, we noted that the RF model is reliable to predict optical bandgaps of lead halide perovskites. The predicted optical bandgap was not directly correlated with the performance of corresponding devices, due to factors limiting the electrical properties and interfacial phenomena. In this context, we anticipate the widening of the descriptor pool with more inputs on the charge transport layers, device architecture, interface properties, crystal size, halide segregations, ion migration, phase stability, and induced losses is detrimental to the success of the ML in predicting the solar energy conversions and this could be further extended to other energy devices such as light-emitting diodes, batteries, photodetectors, etc.

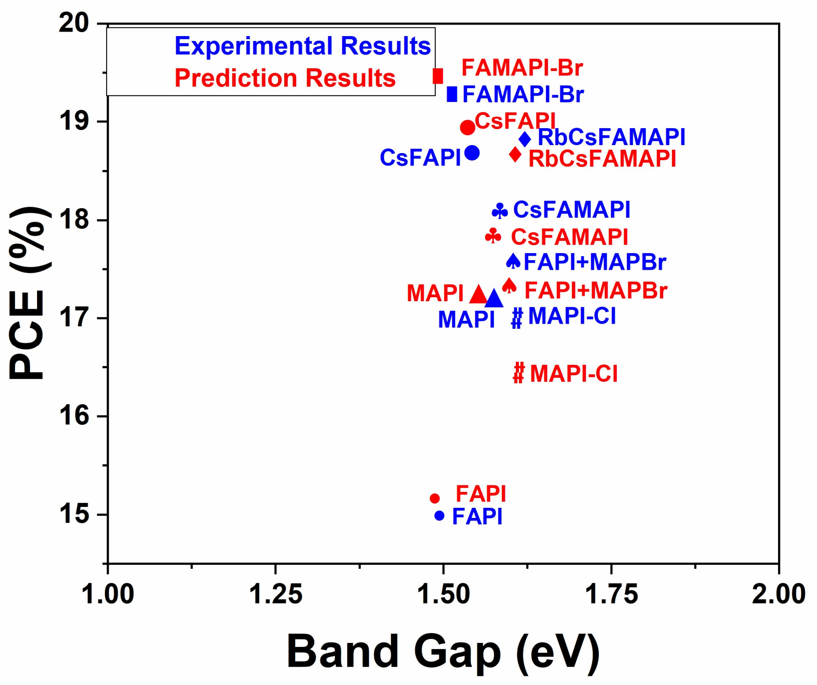

Figure 5. Efficiency vs bandgap energy of various perovskite from the machine learning simulation and experimental data.

4. Conclusions

We developed a random forest model to predict the bandgap of different halide perovskite and their performance evaluation in perovskite solar cells. We investigated the influence of different perovskite compositions on optoelectronic features and photovoltaic performance and validated using the RF model. Our model showed exceptional performance in predicting the optical bandgaps with a high R^2 value of >0.99 and demonstrated that this knowledge can be used to design new lead halide perovskites through accurate bandgap predictions. Further, our random forest model showed judicious fitting of *J-V* curves and predicted the PCEs which is in agreement with the experimental data. This signals the suitability of the employed prediction approach in this work as an effective, reliable, and fast that can be implemented to the variety of materials for solar cell applications, to allow the acceleration of materials discovery and renaissance for rapid screening.

Supplementary Information

Supplemental information includes Figure S1-S10 and Table S1 and can be found with this article online.

Acknowledgments

This work has received funding from the European Union H2020 Programme under the European Research Council Consolidator grant [MOLEMAT, 726360].

Author Contributions

ECG, MOY, EE process the data, made coding and validated for ML approach, AUO supervised and assisted this. MPUH, MP, NH, PH fabricated the devices, collected the experimental data, and made inputs, SK supervised and assisted in all the calculations. S.A conceptualized, supervised, and directed the research. MOY, MPUH, SK, AUO, SA contributed to the first draft and commented on the completion of the final document. ECG, MOY, and MPUH contributed equally to this work.

Conflict of Interests

The authors declare no competing interests.

References

[1] Q. Tao, P. Xu, M. Li, W. Lu, *npj Comput Mater.* 2021, **7**, 1-18.

[2] M.O. Yildirim, E. C. Gok, N. H. Hemasiri, E. Eren, S. Kazim, A. U. Oksuz, S. Ahmad, *ChemPlusChem* 2021*, **86**,* 785–793.

[3] Y. Li, Y. Lu, X. Huo, D. Wei, J. Meng, J. Dong, B. Qiao, S. Zhao, Z. Xu, D. Song*, RSC Adv*. 2021, **11**, 15688-15694.

[4] M. P. U. Haris, S. Kazim, S. Ahmad, *ACS Appl. Energy Mater.* 2021*,***4***,*2600–2606.

[5] J. Li, B. Pradhan, S. Gaur, J. Thomas, *Adv. Energy Mater*. 2019, **9**, 1901891.

[6] A. Gagliardi, A. Abate, *ACS Energy Lett.*2017*,* **3**, 163-169.

[7] E. Mosconi, J. M. Azpiroz, and F. D. Angelis, *Chem. Mater.*2015, **27**, 4885–4892.

[8] W. Kohn, LJ. Sham, *Phys Rev.*1965, **140**, A1133.

[9] X. Yang, L. Li, Q. Tao, W. Lu, M. Li, *Comput. Mater. Sci.* 2021*,* **196**, 110528.

[10] G. Zhang, J. Yuan, Y. Mao, Y. Huang, *Comput. Mater. Sci.,* 2021, **186**, 109998.

[11] V. M. Le Corre, T. S. Sherkar, M. Koopmans, L. J. A.Koster*, Cell Reports Physical Science* 2021*,* **2**, 100346.

[12] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li & J. Wang, *Nat. Commun.* 2018, **9**, 3405.

[13] W. A. Saidi, W. Shadid, I. E. Castelli, *npj Comput. Mater.* 2020, **6**, 36.

[14] K. Takahashi, L. Takahashi, I. Miyazato, Y. Tanaka, *ACS Photonics* 2018, **5**, 771−775.

[15] Z. Weida, Z. Huiran, H. Hongqing, L. Yao, L. Shengzhou, D. Guangtai, Z. Jincang, *Chin. J. Nonfer. Met.* **2019**, 803–807.

[16] P. Huang, Y. Liu, K. Zhang, L. Yuan, D. Li, G. Hou, B. Dong, Y. Zhou, B. Song, Y. Li, *J. Mater. Chem. A 2017,* **5**, 24275–24281.

[17] G. Yang, C. Chen, F. Yao, Z. Chen, Q. Zhang, X. Zheng, J. Ma, H. Lei, P. Qin, L. Xiong, W. Ke, G. Li, Y. Yan, G. Fang, *Adv. Mater.* 2018, **30**, 170602.

[18] F. Hutter, L. Kotthoff, J. Vanschoren, Automated machine learning: methods, systems, challenges, Springer Nature. **2019**

[19] D. Bzdok, N.Altman, M. Krzywinski, *Nat. Methods*. 2018, **15**, 233-234.

[20] A. C. Müller, S. Guido, Introduction to Machine Learning with Python, O'Reilly Media Inc. **2016**

[21] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, The MIT Press, Cambridge, Massachusetts **2016**

[22] J.-W. Lee, D.-H. Kim, H.-S. Kim, S.-W. Seo, S. M. Cho, N.-G.Park, *Adv. Energy Mater.* 2015, **5**, 1501310.

[23] M. Lyu, N.-G. Park, *Sol. RRL* 2020, **4**, 2000331.

[24] N. Arobi, S.M. Amir-Al Zumahi, K. Ibrahim, M. M.r Rahman, M. K. Hossain, M. Mahbubur R. Bhuiyan, H. Kabir, A. Amri, M. A. Hossain, F. Ahmed, *J. Solid State Chem.* 2021, **298**, 122137.