NJC

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d1nj01518d

Prediction of band gap for 2D hybrid organic–inorganic perovskites by using machine learning through molecular graphics descriptors†

Zhongyu Wan, *[ab] Quan-De Wang,*[a] Dongchang Liu[c] and Jinhu Liang[d]

Two-dimensional (2D) hybrid organic–inorganic perovskites (HOIPs) have attracted considerable attention for their promising applications in solar cells and optoelectronics. However, the fast and accurate prediction of the basic band structure of 2D HOIPs is still challenging because the traditional trial-and-error experimental methods or first-principles calculations are usually inefficient. Herein, we introduce machine learning (ML)-aided models with simple descriptors based on molecular graphics and adjacency matrices for the first time to determine the band gap for 2D A_2BX_4 HOIPs, which avoids time-consuming *ab initio* calculations. The multiple stepwise regression algorithm is employed to select 6 important descriptors to represent the electronic and structural features of the molecules. Sixteen competing algorithms including artificial neural network (ANN), regression tree (RT), support vector machine (SVM), Gaussian process regression (GPR) and ensemble of regression tree (ERT) multiple models are used to derive ML models to determine the band gap of 136 2D HOIPs, and the ANN model shows the best accuracy. Based on the developed ANN model, five 2D A_2BX_4 HOIPs with band gaps close to the theoretical Shockley–Queisser limit (1.34 eV) are screened, which are probably excellent candidates for optoelectronics. This work reveals that ML in combination with simple descriptors can serve as an excellent strategy for the fast prediction of the key properties of HOIPs with a high accuracy.

Received 28th March 2021,
Accepted 23rd April 2021

DOI: 10.1039/d1nj01518d

rsc.li/njc

Introduction

Perovskite materials with the general structural formula of ABX_3 have attracted tremendous interest of researchers from both academia and industry due to their great application potential in photovoltaics, optoelectronics, and superconductors.[1–3] As an important subclass, hybrid organic–inorganic perovskites (HOIPs) have attracted significant research attention because of their easy fabrication and structural diversity which are difficult to achieve in purely inorganic perovskites. Specifically, due to longer-lived carriers, a higher absorption coefficient, and more diverse compositions and structures, HOIPs demonstrate great application potential in the field of photovoltaic materials.[4] During the past few years, perovskite-based optoelectronic materials have achieved huge breakthroughs in the field of solar cells, and their photoelectric conversion efficiency has increased from 3.8% to more than 25% currently,[5] and the lifetime of solar cells has also increased to thousands of hours.[6] In addition, due to their special structure, HOIPs also exhibit excellent light-emitting characteristics, which show huge application potential in the field of photodiodes.[7] Traditionally, earlier studies were mainly focused on HOIPs with three-dimensional (3D) structures. Compared with 3D HOIPs, two-dimensional (2D) HOIPs have low dielectric constants because of the organic layer, and the inorganic layer has a high dielectric constant.[6] Thus, 2D HOIPs demonstrate a unique quantum well structure and the dielectric confinement effect, showing excellent application prospects. Hence, 2D HOIPs have attracted phenomenal research interest and several breakthroughs have been made in the research of 2D perovskite materials.[6] But there are still great challenges and difficulties in this field. In particular, it is difficult to achieve 2D HOIPs with band gaps, a key parameter affecting their potential for applications in optoelectronics, close to the theoretical Shockley–Queisser limit (1.34 eV).[8] Thus, the synthesis of 2D HOIPs with diverse structures and the analysis of their optoelectronic properties remain challenging in this field, since the relationship between the structure and properties of 2D HOIPs is not obvious.[9]

[a] *Low Carbon Energy Institute and School of Chemical Engineering, China University of Mining and Technology, Xuzhou, 221008, People's Republic of China. E-mail: quandewang@cumt.edu.cn, zhongyuwanxzit@163.com*
[b] *School of Science, City University of Hong Kong, Hong Kong SAR 999077, People's Republic of China*
[c] *School of Science, Xi'an Polytechnic University, Xi'an, Shaanxi 710048, People's Republic of China*
[d] *School of Environment and Safety Engineering, North University of China, Taiyuan 030051, People's Republic of China*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d1nj01518d

The structural diversity of different functional groups in the ABX_3 structure makes experimental screening and synthesis of 2D HOIPs for specific properties challenging. Generally, accurate prediction of band gap requires first-principles density functional theory (DFT) calculations, which is very expensive in terms of time and cost for comprehensive search in large compositional and structural spaces based on traditional trial-and-error experiments. In recent years, high-throughput computational materials design has become an effective and popular method for novel materials discovery and has been applied in many fields.[8] This method requires large-scale databases either from existing experiments or from first-principles calculations and selects appropriate material descriptors for target applications. After that, machine learning (ML) or related artificial intelligence (AI) methods can be used to build prediction models.[10–18] For example, Wang *et al.* used 30 descriptors such as the tolerance factor, the octahedron factor, metal electronegativity and organic molecular polarizability to discover new 3D perovskite materials.[19] Morgan *et al.* used ion radius, bond length, and orbital radius as descriptors for the prediction of the thermodynamic stability of perovskite oxides.[20] All these methods use descriptors from quantum chemistry or first-principles calculations, which is difficult for experimental researchers and is also computationally demanding in the construction of databases. In addition, the selected descriptors are needed to be computationally viable and should be able to accurately describe the desired materials properties for target applications. Therefore, physically meaningful and simple descriptors that can be easily computed are the key to develop efficient ML methods for materials discovery.

Based on the above considerations, this work introduces an efficient ML-aided model with simple descriptors for the first time to accurately predict the band gap of 2D A_2BX_4 HOIPs. The introduced descriptors are fully based on the molecular structure information characterized by molecular graphics and adjacency matrices, and can be computed within 1 second for the considered 2D A_2BX_4 HOIPs. To reduce the complexity of the established ML model, a multiple stepwise regression algorithm is employed to select important descriptors. A total of sixteen competing algorithms are then used to derive ML models to determine the band gap of 136 2D HOIPs. The developed ML-aided model can quickly and effectively predict the band gap of 2D HOIPs and provide guidance for the design of new materials with excellent optical properties.

Results and discussion

The dataset

Most of the materials have a wide band gap, which results in a narrow light absorption range of photovoltaic devices, which in turn leads to a decrease in the current density of the solar cell. Therefore, band gap is the most important indicator that determines the performance of a perovskite solar cell.[21] This work focuses on HOIPs with the general structural formula of A_2BX_4, which is the most common structure in two-dimensional single-layer Ruddlesden–Popper phase perovskites.[22] The organic layer and the inorganic layer are connected by interactions to form a stable structure.[23,24] The A site is a +1 valent organic cation, the B site is a +2 valent metal ion, and the X site is a −1 valent halide ion. The different spatial structures even for HOIPs with the same compositions still have different band gaps due to the different lattice structures and parameters. Thus, the spatial structure effect is still needed to be considered. For this purpose, an effective and comprehensive dataset is critical for ML model training. This work employs the 2D perovskite database built by Marchenko *et al.*,[25] which includes high-quality values of the band gap, the structural formula of each site ion, and the crystal structure. The detailed structural information of the selected 136 2D HOIPs is provided in the ESI,† and Fig. 1 shows the general information of the studied 2D HOIPs.

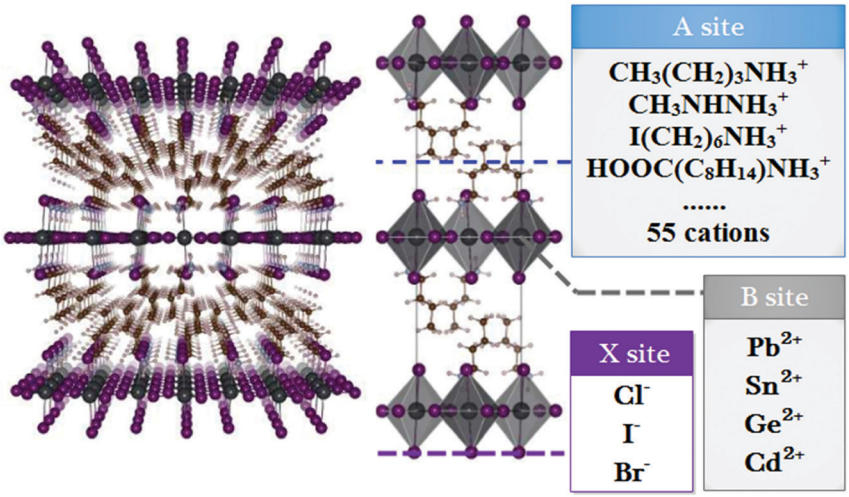


Fig. 1 The composition of 136 2D HOIPs, which include 55 A-site ions (organic ions), 4 B-site ions (metal ions) and 3 X-site ions (halide ions).

Computational work

Selection of molecular descriptors. A workflow for model training is depicted in Fig. 2. Generally, molecular descriptors can well characterize the structure of a molecule or an ion, and express the ion structure in the numerical form.[25,26] However, due to the complexity of first-principles and quantum chemical calculations on 2D HOIPs, simple descriptors that can be quickly estimated are critical for these materials. For this purpose, the PaDEL-descriptors developed by Yap *et al.*[27] are introduced in this work since they can be quickly estimated and have physical meanings. Specifically, for the A-site ion, a total of 634 1D and 2D descriptors based on based on molecular graphics and adjacency matrices are selected. For the B- and X-site ions, their atomic weight, the number of extra-nuclear electrons and the number of protons are selected as descriptors. For a given ion, these descriptors are all known constants. Details of the computational methods of descriptors can be found in ref. 28–32. To characterize the crystal structure of 2D HOIPs, the lattice constants are used to describe perovskite crystals with different spatial structures. Considering the length of the three sides (a, b, c) and the angles of the three sides (α, β, γ), three descriptors, namely, $\Delta\alpha = |90^\circ - \alpha|$, $\Delta\beta = |90^\circ - \beta|$ and $\Delta\gamma = |90^\circ - \gamma|$, are defined, and a total of 649 descriptors are obtained to characterize the properties of a 2D-HOIP molecule. Although these descriptors can be quickly computed, the selection of a small number of suitable descriptors can significantly improve the computational efficiency in ML model training and its applications. To find the most suitable and relevant descriptors for ML model training, multiple stepwise regression (MSR) with a fast computational speed is used for the rapid screening of a large number of descriptors.[33] The MSR method analyzes the relationship between the independent variable (descriptor) and the dependent variable (band gap) to obtain a subset of independent variables with the greatest degree of correlation. The degree of correlation between the band gap and descriptors is quantitatively described by the coefficient of determination (R^2), and the leave-one-out method (LOO) is used for the cross-validation of the model. In addition, to avoid overlap among the selected important features and ensure that there is no multicollinearity, the variance inflation factor (VIF) is introduced to control the independence of the selected descriptors. To summarize, the band gaps of 136 HOIPs in the entire data set are used as dependent variables and 649 descriptors are used as independent variables. Details of the MSR analysis results are shown in the ESI.† It is demonstrated that the values of R^2 and cross-validation coefficient (Q^2) increase with the increasing in the number of important molecular descriptors, which results in an improvement in the fitting ability and stability of the model. However, the VIF values are larger than 10 when the number of descriptors exceeds 6, which exhibits collinearity problems. Thus, the MSR result with 6 variables is selected as the important feature for further ML model training. The corresponding descriptors are the atomic weight of the ion at the X site (M_X), the atomic weight of the B-site ion (M_B), Moran's autocorrelation weighted by charges (MATS4c), Crippen's LogP that represents the lipophilicity and hydrophilicity of a substance, the newly defined descriptor ($\Delta\gamma$) representing the absolute value of deviation from 90°, and Geary's autocorrelation weighted by Sanderson's electronegativities (GATS4e). According to the results of the MSR, the determination coefficient of the 6 descriptors and the band gap in the model reached 0.931, indicating that the correlation between these 6 descriptors and the band gap reached 93.1%. In addition, the obtained relationship reveals that the band gap is negatively related to M_X, MATS4c and GATS4e, while it is positively related to the other three descriptors. Generally, the smaller the band gap, the greater the current density, and the better the performance of optical materials for perovskites. The derived descriptors are in fact physically meaningful, and this information first provides a qualitative method for materials design. For example, the atomic weight of the ion at the X site is negatively correlated with the band gap, which shows that choosing an ion with a larger atomic weight can increase the band gap of HOIPs, while the atomic weight of the ion at the B site exhibits the opposite effect. MATS4c is a descriptor related to charge distribution and GATS4e is a descriptor related to electronegativity, which indicates that the band gap value is related to the charge distribution of the ions at the A site. Based on the calculation results for 2D HOIPs (No. 9–16 and No. 48–66 shown in Table S2, ESI†), as the number of $-CH_2-$ groups decreases, the values of the two descriptors increase, indicating

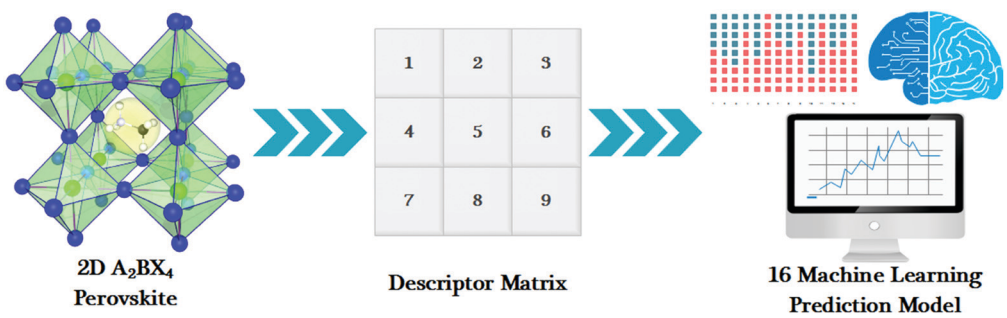


Fig. 2 Predicting the band gap of HOIPs based on machine learning methods. The methods used include artificial neural networks, 6 types of support vector machine, 3 types of regression tree, 2 types of ensemble of regression tree and 4 types of Gaussian process regression.

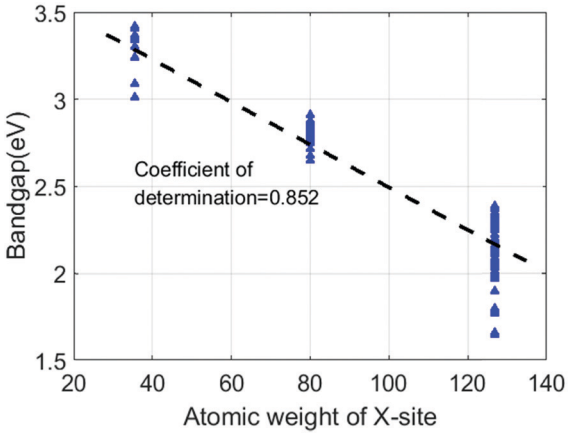


Fig. 3 Correlation diagram between the atomic weight of the X-site ion and the band gap of 2D HOIPs.

that the charge distribution on the A-site ion is more concentrated on the electronegative atoms. Thus, the concentration of electric charges increases the interactions of the organic ions at the A site, which makes the organic layer and the inorganic layer more tightly combined and further reduces the band gap value. The same conclusion can be derived from the descriptor Crippen's LogP. The descriptor representing the value of LogP, which represents the lipophilicity and hydrophilicity of a substance, is related to the electronegativity of the group in the A-site ion. Therefore, reducing the length of the molecular chain at the A site or connecting a strongly electronegative atom at the end of the ion chain may help to reduce the band gap. It is worth noting that, according to the MSR results of the first model only involving the M_X descriptor as shown in Fig. 3, the correlation between M_X and the energy gap reaches 86.8%, indicating that the atomic weight of the X-site ion is strongly correlated with the band gap of 2D HOIPs. Thus, the choice of X-site ions greatly determines the range of the band gap.

Machine learning models

To obtain an optimal ML model for the design of 2D HOIPs, sixteen competing ML algorithms, as shown in Fig. 2, including artificial neural network (ANN), regression tree (RT), support vector machine (SVM), Gaussian process regression (GPR) and ensemble of regression tree (ERT) multiple models, are screened and systematically compared. The entire dataset is divided into a training set with 109 HOIPs (80%) and a test set with 27 HOIPs (20%). The training set is used to train the ML model to select the model with the best fitting performance. The R^2 values are used to evaluate their ability to fit the data. The test set is used to verify the external predictive ability of the obtained ML model and also to check whether the model overcomes the problem of overfitting. Details of the parameters used in the 16 ML models can be found in Tables S3–S6 in the ESI.† In the ANN model, the most critical parameter is the number of hidden layers. Based on Andrea's and Xu Lu's rule,[33,34] the number of hidden layer nodes is selected as 9. All the data are then normalized, and the 6 screened descriptors are used as the input for the training of the 16 ML models. The coefficients of determination, R^2, corresponding to these models can be seen in Fig. 4. The determination coefficient of the ANN model is 0.99, indicating that this model exhibits an excellent fitting ability to the dataset. The R^2 of the test set is 0.98, indicating that the model has excellent fitting and external predictive capabilities. Moreover, the difference between the coefficient of determination of the training set and that of the test set is only 0.01, indicating that the model does not have the problem of overfitting during the training process. The performance of the test set of 16 ML models can be seen in Table 1, including the mean absolute error (MAE), root mean square error (RMSE) and mean square error (MSE).

Validation of the ANN model

In order to verify the prediction accuracy of the ANN model, in this work, we use the experimental values of the band gap to

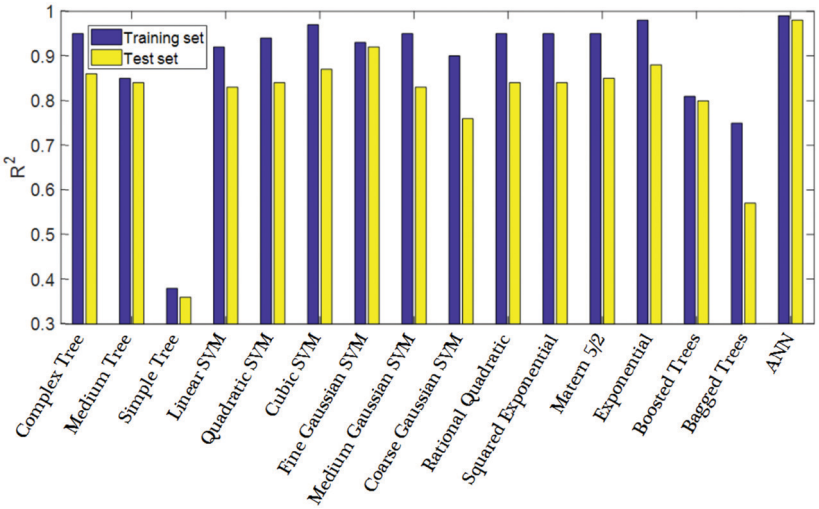
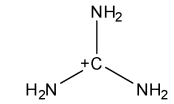
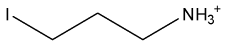
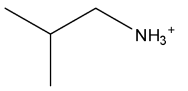
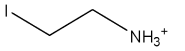
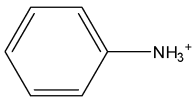


Fig. 4 Coefficient of determination of 16 ML models.

Table 1 Performance of the test set of 16 ML models

Prediction model		RMSE	MSE	MAE
RT	Complex tree	0.17	0.03	0.14
	Medium tree	0.19	0.04	0.16
	Simple tree	0.41	0.16	0.45
SVM	Linear SVM	0.19	0.04	0.16
	Quadratic SVM	0.19	0.04	0.15
	Cubic SVM	0.17	0.03	0.14
	Fine Gaussian SVM	0.14	0.02	0.1
	Medium Gaussian SVM	0.19	0.04	0.15
	Coarse Gaussian SVM	0.23	0.05	0.19
GPR	Rational quadratic	0.19	0.04	0.16
	Squared exponential	0.19	0.04	0.16
	Matern 5/2 GPR	0.18	0.03	0.15
	Exponential GPR	0.17	0.03	0.13
ET	Boosted trees	0.21	0.04	0.16
	Bagged trees	0.31	0.1	0.22
ANN	ANN	0.09	0.01	0.07

Table 2 Predicted 2D HOIPs with high band gaps

No.	A site	B site	X site	Band gap (eV)
1		Pb^{2+}	I^-	1.48
2		Cd^{2+}	I^-	1.45
3		Cd^{2+}	I^-	1.48
4		Cd^{2+}	Cl^-	1.47
5		Ge^{2+}	I^-	1.50

compare the predicted values. The name and the experimental and predicted values of 2D HOIPs used for verification can be seen in Table S8 (ESI†). As shown in Fig. 5, most of the scattered points are concentrated around the straight line $y = x$, and there are no obvious outliers. This shows that ANNs exhibit excellent predictive performance, and the predicted value is very close to the experimental value.

Applications of the ANN model

The ANN model is employed for the prediction of materials for optoelectronic applications based on band gap. A dataset with 660 different structural 2D A_2BX_4 HOIPs is obtained by the combination of 55 A sites, 4 B sites, and 3 X sites. The γ angle of nearly all the molecules is 90°. By neglecting the known 136 2D HOIPs used for the training and verification of the ANN model, the band gap of the remaining 524 2D A_2BX_4 HOIPs is predicted using the developed ANN model. Based on the ANN model, Table 2 lists the 5 2D HOIPs with band gaps close to the limit value; four of the five 2D HOIPs we selected do not contain lead. The lead-free perovskite has a band gap value close to 1.34 eV and overcomes the toxicity of lead, and is expected to become an excellent optoelectronic material in the future.

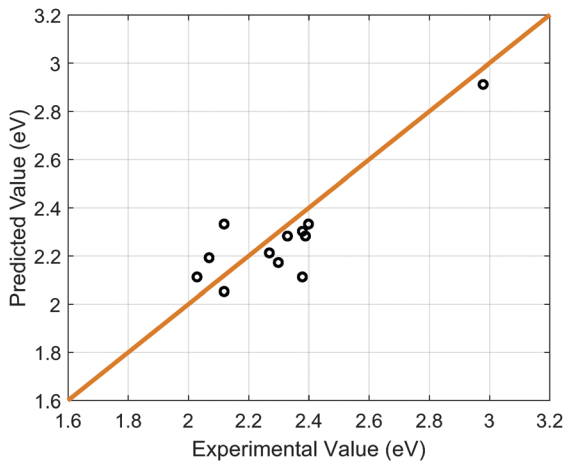


Fig. 5 Scatter plots of predicted and experimental values.

Conclusion

In this work, we successfully developed a ML model with simple descriptors to predict the band gap for 2D HOIPs. The descriptors can be quickly and easily computed for the given molecular compositions of 2D HOIPs based on molecular graphics and adjacency matrices. To characterize the crystal structure of 2D HOIPs, three descriptors about angular deviation are defined. MSR analysis is used to derive 6 important features that do not have collinearity, which are further used for the training of 16 ML models. By comparing the value of the determination coefficient, the ANN model shows the best performance, and its training set has $R^2 = 0.99$ and test set has $R^2 = 0.98$. An in-depth analysis based on the results of the MSR model is performed, and the physical meanings of the descriptors are discussed, which provide efficient qualitative information for materials design. The ANN model is employed to predict 2D A_2BX_4 HOIPs with large band gaps and five potential 2D A_2BX_4 HOIPs with excellent optoelectronic properties. This work provides an effective and computationally efficient framework for a ML model developed for materials design.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities of China (No. 2020ZDPYMS05).

References

1 M. A. Green, A. Ho-Baillie and H. J. Snaith, The emergence of perovskite solar cells, *Nat. Photonics*, 2014, **8**, 506–514.

2 M. H. Kumar, N. Mathews and P. P. Boix, Decoupling light absorption and charge transport properties in near IR-sensitized Fe_2O_3 regenerative cells, *Energy Environ. Sci.*, 2013, **6**, 3280–3285.
3 B. Saparov and D. B. Mitzi, Organic–Inorganic Perovskites: Structural Versatility for Functional Materials Design, *Chem. Rev.*, 2016, **116**, 4558–4596.
4 A. Polman, M. Knight and E. C. Garnett, Photovoltaic materials: present efficiencies and future challenges, *Science*, 2016, **352**, 4424.
5 P. Schulze, A. J. Bett and M. Bivour, 25.1% High Efficient Monolithic Perovskite Silicon Tandem Solar Cell with a High Band Gap Perovskite Absorber, *Sol. RRL*, 2020, **4**, 2000152.
6 D. Zhao, Y. Yu and C. Wang, Low-bandgap mixed tin–lead iodide perovskite absorbers with long carrier lifetimes for all-perovskite tandem solar cells, *Nat. Energy*, 2017, **2**, 17018.
7 G. Grinblat, Y. Li and M. P. Nielsen, Efficient third harmonic generation and nonlinear subwavelength imaging at a higher-order anapole mode in a single germanium nanodisk, *ACS Nano*, 2016, **11**, 953–960.
8 N. A. Astani, F. Jahanbakhshi and M. Mladenovic, Ruddlesden-Popper Phases of Methylammonium-based 2D Perovskites with 5-Ammonium Valeric Acid AVA2MAn-1Pb*n*I3*n*+1 with n = 1, 2 and 3, *J. Phys. Chem. Lett.*, 2019, **10**, 3543–3549.
9 W. Shockley and H. J. Queisser, Detailed Balance Limit of Efficiency of p–n Junction Solar Cells, *J. Appl. Phys.*, 1961, **32**, 510–519.
10 Y. Zhang and X. Xu, Predicting lattice parameters for orthorhombic distorted-perovskite oxides via machine learning, *Solid State Sci.*, 2021, **113**, 106541.
11 H. Park, R. Mall, A. Ali, S. Sanvito, H. Bensmail and F. El-Mellouhi, Importance of structural deformation features in the prediction of hybrid perovskite bandgaps, Comp, *Mater. Sci.*, 2020, **184**, 109858.
12 Y. Zhang and X. Xu, Machine learning glass transition temperature of polyacrylamides using quantum chemical descriptors, *Polym. Chem.*, 2021, **12**, 843–851.
13 Y. Zhang and X. Xu, Machine learning lattice constants for cubic perovskite A22 + BB′O6 compounds, *CrystEngComm*, 2020, **22**, 6385–6397.
14 Y. Zhang and X. Xu, Machine learning lattice parameters of monoclinic double perovskites, *Int. J. Quantum Chem.*, 2021, **121**, e26480.
15 Y. Zhang and X. Xu, Machine Learning Band Gaps of Doped-TiO2 Photocatalysts from Structural and Morphological Parameters, *ACS Omega*, 2020, **5**, 15344–15352.
16 Y. Sheng, Y. Wu, J. Yang, W. Lu, P. Villars and W. Zhang, Active learning for the power factor prediction in diamond-like thermoelectric materials, *npj Comput. Mater.*, 2020, **6**, 1–7.
17 A. Chen, X. Zhang and Z. Zhou, Machine learning: accelerating materials development for energy storage and conversion, *InfoMat*, 2020, **2**, 553–576.
18 Z. Zhou, Journal of Materials Chemistry A and Materials Advances Editor's choice web collection: machine learning for materials innovation, *J. Mater. Chem. A*, 2021, **9**, 1295–1296.
19 Y. Li and K. Yang, High-throughput computational design of halide perovskites and beyond for optoelectronics, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2020, **11**, e1500.
20 S. H. Lu, Q. H. Zhou, Y. X. Ouyang, Y. Guo, Q. Li and J. L. Wang, Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning, *Nat. Commun.*, 2018, **9**, 3405.
21 W. Li, R. Jacobs and D. Morgan, Predicting the thermodynamic stability of perovskite oxides using machine learning models, *Comput. Mater. Sci.*, 2018, **150**, 454–463.
22 T. Félix, V. C. David and C. Quarti, Phonon coherences reveal the polaronic character of excitons in two-dimensional lead halide perovskites, *Nat. Mater.*, 2019, **18**, 406.
23 F. Haque, S. Lim and M. Mativenga, Ambient-Air-Processed Ambipolar Perovskite Phototransistor With High Photodetectivity, *IEEE Trans. Electron Devices*, 2020, **99**, 1–6.
24 S. A. Dar, M. Ali and V. Srivastava, Investigation on bismuth-based oxide perovskites MBiO3 (M = Rb, Cs, Tl) for structural, electronic, mechanical and thermal properties, *Eur. Phys. J. B*, 2020, **93**, 102.
25 E. I. Marchenko, S. A. Fateev and A. A. Petrov, Database of 2D hybrid perovskite materials: open-access collection of crystal structures, band gaps and atomic partial charges predicted by machine learning, *Chem. Mater.*, 2020, **32**, 7383–7388.
26 W. J. Dunn, A. J. Hopfinger and C. Catana, Solution of the conformation and alignment tensors for the binding of trimethoprim and its analogs to dihydrofolate reductase: 3D-quantitative structure-activity relationship study using molecular shape analysis, 3-way partial least-squares regression, and 3-way factor analysis, *J. Med. Chem.*, 1996, **39**, 4825.
27 C. W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, *J. Comput. Chem.*, 2011, **32**, 1466–1474.
28 L. B. Kier, L. H. Hall and J. W. Frazer, An index of electro-topological state for atoms in molecules, *J. Math. Chem.*, 1991, **7**, 229–241.
29 T. A. Samad, A. Rebbapragada, E. Bell, Y. Zhang, Y. Sidis, S. J. Jeong, J. A. Campagna, S. Perusini, D. A. Fabrizio, A. L. Schneyer, H. Y. Lin, A. H. Brivanlou, L. Attisano and C. J. Woolf, DRAGON, a Bone Morphogenetic Protein Co-receptor, *J. Biol. Chem.*, 2005, **280**, 14122–14129.
30 K. Roy and G. Ghosh, QSTR with Extended Topochemical Atom Indices. 2. Fish Toxicity of Substituted Benzenes, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 559–567.
31 K. Roy and R. N. Das, On some novel extended topochemical atom (ETA) parameters for effective encoding of chemical

information and modeling of fundamental physicochemical properties, *SAR QSAR Environ. Res.*, 2011, **22**, 451–472.
32 R. Nilakantan, D. S. Nunn, L. Greenblatt, G. Walker, K. Haraki and D. Mobilio, A family of ring system-based structural fragments for use in structure-activity studies: database mining and recursive partitioning, *J. Chem. Inf. Model.*, 2006, **46**, 1069–1077.
33 L. Xu, Y. Wu and C. Hu, A QSAR of the toxicity of aminobenzenes and their structures, *Sci. China, Ser. B: Chem.*, 2000, **9**, 1–4.
34 T. A. Andrea and H. Kalayeh, Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors, *J. Med. Chem.*, 1991, **34**, 2824–2836.