Machine-Learning Prediction of the Computed Band Gaps of Double Perovskite Materials

Junfei Zhang[1], Yueqi Li[2], and Xinbo Zhou[3]

[1] School of Computing and Information Systems, The University of Melbourne, Melbourne, Victoria, Australia
[2] College of Physical Science and Technology, Xiamen University, Xiamen, Fujian, China
[3] Faculty of Information Technology, Beijing University of Technology, Beijing, China

* Emails: junfei.zhang@student.unimelb.edu.au; liyueqi@stu.xmu.edu.cn; xbxb2020@126.com

Abstract

Conventional electronic structure methods based on density functional theory (DFT) suffer from not only high computational cost that scales cubically with the number of electrons, but also limited accuracy arising from the approximations of the exchange-correlation functional. In particular, the latter issue is particularly severe for insulators and semiconductors, whose band gaps are systematically underestimated by DFT. Surrogate methods based on machine learning has garnered much attention as a viable alternative to bypass these limitations, especially in prediction of solid-state band gaps. Here, we construct a random forest regression model for band gaps of double perovskite materials, using a dataset of 1306 band gaps computed with the GLLBSC (Gritsenko, van Leeuwen, van Lenthe, and Baerends solid correlation) functional. Among the 20 physical features employed, we find that the bulk modulus, superconductivity temperature, and cation electronegativity exhibit the highest importance scores, consistent with the physics of the underlying electronic structure. Using the top 10 features, a model accuracy of 85.6% with a root mean square error of 0.64 eV is obtained, comparable to a previous study employing kernel ridge regression. Our results attest to the potential of machine learning regressions for rapid screening of promising candidate functional materials.

Introduction

Double perovskites ($AA'BB'X_6$) have double the unit cell of single perovskites (ABX_3) with chemically distinct A/A' and B/B' sites[1]. A variety of physical and chemical properties can be engineered by doping the cations with species of different valence states or radii[2]. Due to their stable crystal structure, unique electromagnetic properties, and high catalytic activities, these compounds have much potential as functional materials for environmental protection[3], chemical industry[4], photovoltaics[5], and catalysis[6]. In this regard, optimization and engineering of these properties requires proper description of the underlying electronic structure[7].

In quantum mechanics, the energy of bound electrons becomes quantized,[8] and electrons at the ground state can be excited to higher energy levels by absorbing photons with the corresponding wavelengths. In solid structures, the superposed electronic states form continuous energy bands. In insulators and semiconductors, the band gap is the energy gap across the valence and conduction band where electrons are forbidden to occupy. The magnitude of the band gap plays an important role in many functional materials, such as transistors, photovoltaics, light-emitting diodes, and sensors[9]. For instance, optoelectronic materials are generally wide-band gap semiconductors, while thermoelectric materials are narrow-band gap semiconductors[10]. Hence, accurate and efficient prediction of band gaps of solid materials is crucial for design and engineering of new devices.

One of the most widely used electronic structure methods for evaluating band gaps is density functional theory (DFT)[11]. In the Kohn-Sham formalism[12], the multielectron wavefunction is replaced by fictitious noninteracting states that give rise to the true electron density[13], which enables iterative solution of the single-particle Hamiltonian. However, the exchange-correlation energy, which contains all the quantum mechanical interactions of the electrons, does not have an exact expression in terms of the electron density and as such requires an approximation, such as the local density approximation (LDA)[14] or the generalized gradient approximation (GGA)[15]. Such approximation has limited accuracy, most notably underestimation of the band gap of semiconductors and insulators[16]. Various approaches have been proposed to address this limitation, such as the on-site Hubbard U correction[17], hybrid functionals using fractional exact exchange[18], and quasiparticle methods such as the GW approximation[19]. However, these methods do not always guarantee accurate description of the system, and they can be much more computationally expensive than conventional DFT[20].

Alternative strategy for band gap prediction is machine learning. For example, a support vector regression model was constructed for inorganic solids using experimentally measured band gaps[21], thereby bypassing the limitations of DFT. Another study trained a kernel ridge regression model[22] using band gaps computed with the GLLBSC (Gritsenko, van Leeuwen, van Lenthe, and Baerends solid correlation) functional[23], which demonstrated reasonable agreement with experimental values. These studies attest to the potential of machine learning methods, provided that robust datasets are available for training[24].

Previous studies have shown that random forest regression is well-suited in capturing nonlinearity, as seen across the band gap and the extracted physical features such as the highest occupied energy level[25]. As such, we construct a random forest regression model[26] for predicting the band gap of double perovskite compounds. We build upon a previous kernel ridge regression study[22] employing a dataset of GLLBSC-computed band gaps of 1306 double perovskites. We find that the bulk modulus, superconductivity temperature, and cation electronegativity exhibit the highest importance scores among the 20 physical descriptors employed, consistent with the physics of the underlying electronic structure. A model accuracy of 85.6% with a root mean square error of 0.64 eV is obtained using the top 10 features, comparable to previous studies[8].

METHODS

Random Forest Regression

Random forest regression is a regression method that utilizes multiple decision trees[27], which are constructed by a simple supervised algorithm consisting of a series of if-then-else statements. The randomness is manifested through random sampling of data subsets or random selection of features. Multiple uncorrelated decision trees construct a random forest, where all trees are granted free growth without any pruning. The random forest algorithm can be employed for both classification and regression. For classification, the result is the outcome with the highest turnout among all trees; for regression, the forest takes the average of all trees. The steps to generate a random forest are as follows[26]:

1. From a sample with capacity *N*, randomly select *N* times, each time selecting one data. The resulting *N* samples are used as the node samples of decision trees.
2. Choose a constant *m* smaller than the dataset feature number *M*.
3. When splitting each decision tree, select *m* features from the original *M* features, choosing one feature as the splitting feature of the node.
4. Repeat step 2 for each node until no splitting can occur, i.e. when the next feature is used by the parent node in the last splitting. The tree is always left unpruned to ensure free growth.
5. Repeat steps 1-4 to construct a large amount of decision trees to generate a random forest.

Random forest can manage data with a high dimension of features without performing dimension reduction or feature selection. This is beneficial for the dataset of this study, which involves multiple atomic descriptors of double perovskites. The mutual effects of different features and their significance are also quantified. Although random forest regression has been shown to be computationally efficient and accurate when using a large number of generated trees, the risk of overfitting still exists for data with a large noise. We perform random forest regression as implemented in *scikit-learn*, using the *double_perovskites_gap* dataset available in the *matminer* package[28].

Features

20 atomic features are obtained from the *periodic_table* and *composition* modules of *Pymatgen*[29] (Python Materials Genomics) package:

Average electronegativity
Average cation electronegativity
Average atomic radius
Average van der walls radius
Average Mendeleev number
Average electrical resistivity
Average molar volume
Average thermal conductivity
Average boiling point
Average melting point
Average critical temperature
Average superconduction temperature
Average bulk modulus
Average youngs modulus
Average Brinell hardness
Average rigidity modulus
Average mineral hardness
Average Vickers hardness
Average density of solid phase

Average first ionization energy

The dataset is first converted into a data frame, which is then processed by applying the chemical composition of each compound to corresponding classes and functions in the *Pymatgen* package to obtain the 20 features. Compositional averages are taken for atomic features of a given compound, whereas molecular features are used directly. Missing values are not counted in the calculation of the average.

Results and Discussion

Model selection

Random forest regression has two parameters to be optimized: the number of estimators (*n_estimator*) referring to the number of trees to be built before taking the maximum voting or averages of predictions; and the random seed (*random_state*) for the random generator. Both the accuracy and the computational cost of the model increase with the number of estimators[30]. The cost scales as $O\left(n_{\text{tree}} * m_{\text{try}} * n\log(n)\right)$, where n_{tree} is the number of estimators, m_{try} is the number of variables to sample at each node, and n is the number of records[31]. As such, an optimal number of estimators is needed to ensure a satisfactory model performance.

As shown in **Fig. 1**, the model accuracy reaches a maximum at around 700 estimators and decreases afterwards, which is attributed to overfitting[32]. As such, *n_estimator* is set to 700. On the other hand, the random seed determines the random sampling for the train-test split and may subtly affect the accuracy due to the randomization of the training pipeline. An optimal *random_state* value of 14 is selected.

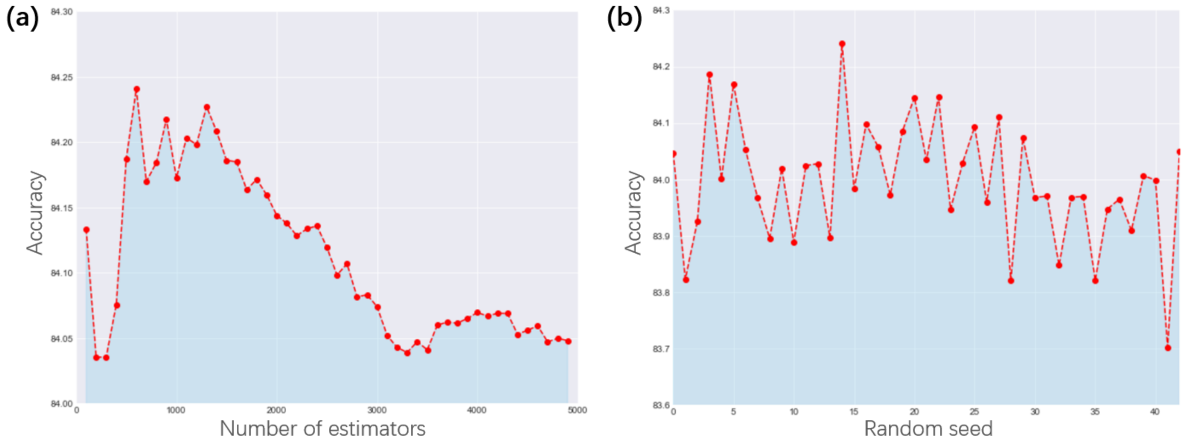

Figure 1. The accuracy of the random forest regression model as a function of (**a**) the number of estimators and (**b**) the random seed, using all 20 physical descriptors.

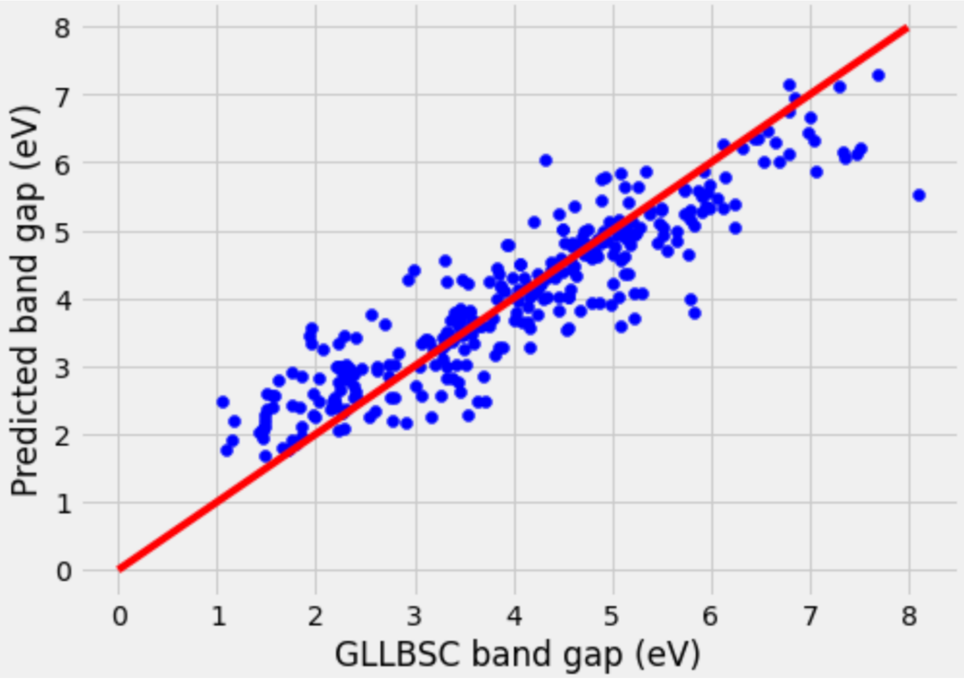

Figure 2. Parity plot of the predicted vs. GLLBSC-computed band gaps, obtained using all 20 physical descriptors and a test/training ratio of 25/75. The parity line is shown in red.

The corresponding parity plot of the model prediction is shown in **Fig. 2**. Using a test/training ratio of 0.25 and all 20 physical descriptors, the model accuracy is 85.1% with a mean absolute error (MAE) of 0.47 eV, a root mean squared error (RMSE) of 0.62 eV, which is comparable to the RMSE value of 0.5 eV reported in a previous kernel ridge regression study of the same dataset.

Feature selection

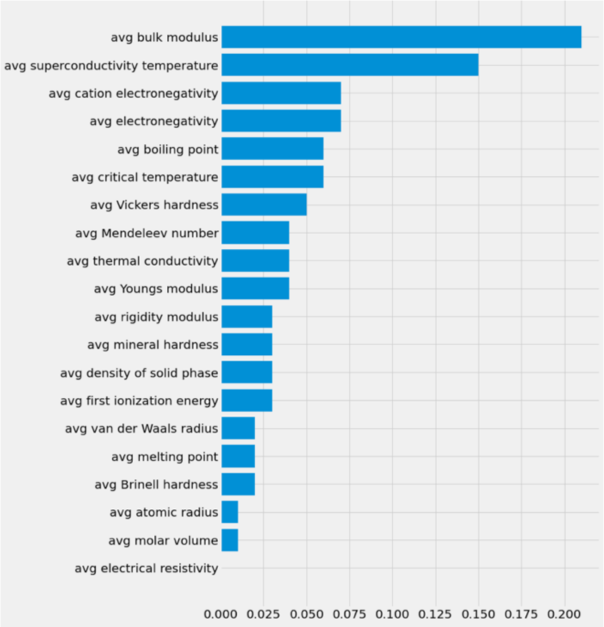

Figure 3. Feature importance of all 20 physical descriptors, obtained from a test/training ratio of 25/75.

The top three features with the highest importance scores are average bulk modulus, superconductivity temperature, and cation electronegativity:

1) Bulk modulus quantifies the elastic property of a solid or fluid under pressure, specifically its resistance to compression[33]. Microscopically, bulk modulus depends on the compressibility of atoms, which affects the extent of the overlap of valence atomic orbitals, and therefore the band gap of the material[34].

2) Superconductivity is the state of matter with no electrical resistance and magnetic penetrability[35]. Given that the magnitude of the band gap determines the electrical conductivity, a material with a relatively small band gap is expected to more easily achieve a superconducting state[36].

3) Electronegativity quantifies the ability of an atom to attract an electron pair in a chemical bond[37]. The cation electronegativity here refers to the electronegativity difference between the oxygen anions and the metal cations. Larger elemental electronegativity difference leads to a larger degree of electron localization around the more electronegative element, which makes it harder for electrons to leap to the conduction band[38].

The low importance scores of some features, such as average electrical resistivity and molar volume, indicates that the dataset contains a large amount of noise, which necessitates feature selection. **Table 1** summarizes the model performance using different number of top features. The performance remains optimal up to top 10 features, which yields an accuracy of 85.6% with RMSE of 0.64 eV. Given the marginal difference in accuracy using 20, 15, and 10 top features, the remainder of the study employs the top 10 features only.

Table 1. Model performance obtained using different number of features with the highest feature importance scores (MAE = mean absolute error; RMSE = root mean squared error; NRMSE = normalized RMSE).

Number of top features	20	15	10	5	3	1
Accuracy (%)	85.1	85.5	85.6	82.3	82.4	65.2
MAE (eV)	0.47	0.46	0.46	0.56	0.57	1.12
RMSE (eV)	0.62	0.62	0.64	0.79	0.81	1.43
NRMSE	0.08	0.07	0.08	0.10	0.10	0.17

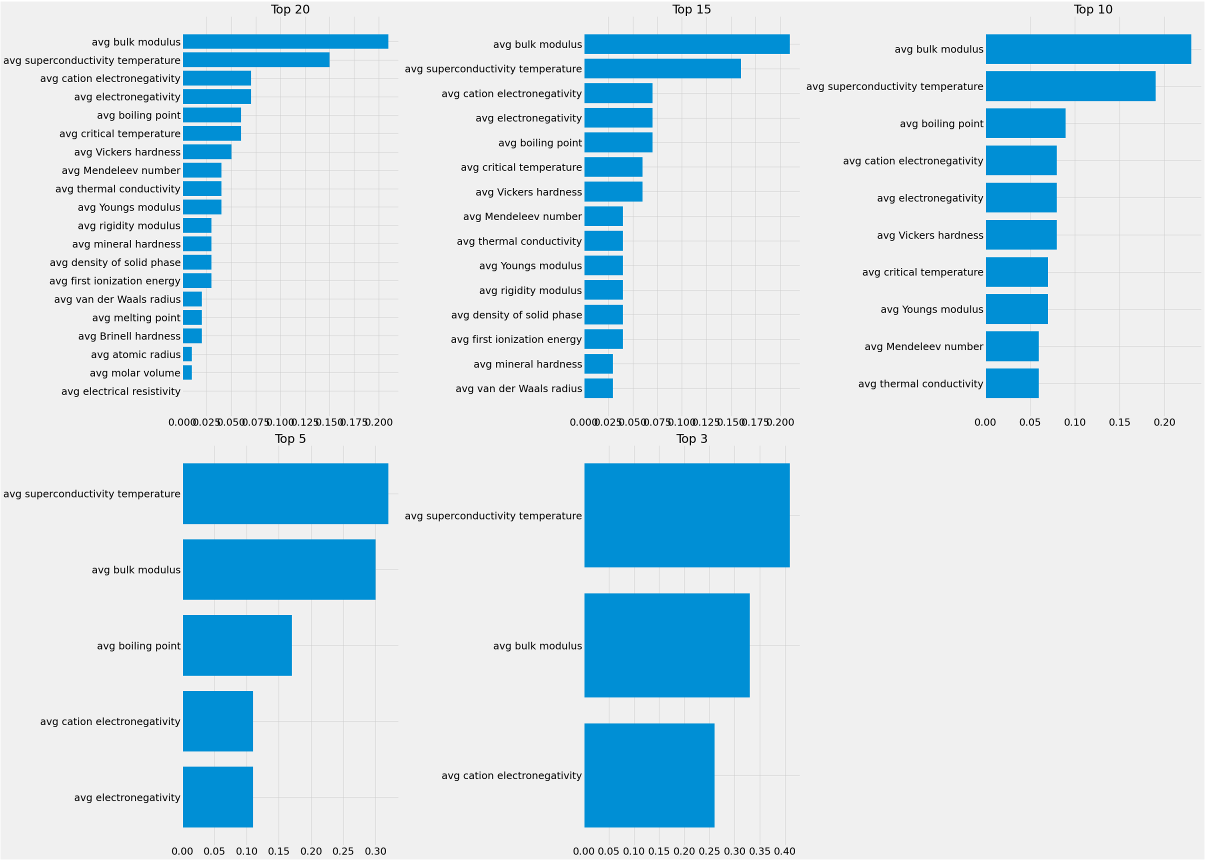

Figure 4. Feature importance scores for models constructed using different number of features with the highest importance scores.

The corresponding importance scores and parity plots are shown in **Figs. 4 & 5**, respectively. The model constructed using the top 10 features exhibits the least deviation of the data points from the parity line. Moreover, the models overall have a tendency to show larger underestimation for larger band gap values, which can potentially be attributed to the limited accuracy of the GLLBSC functional itself[39].

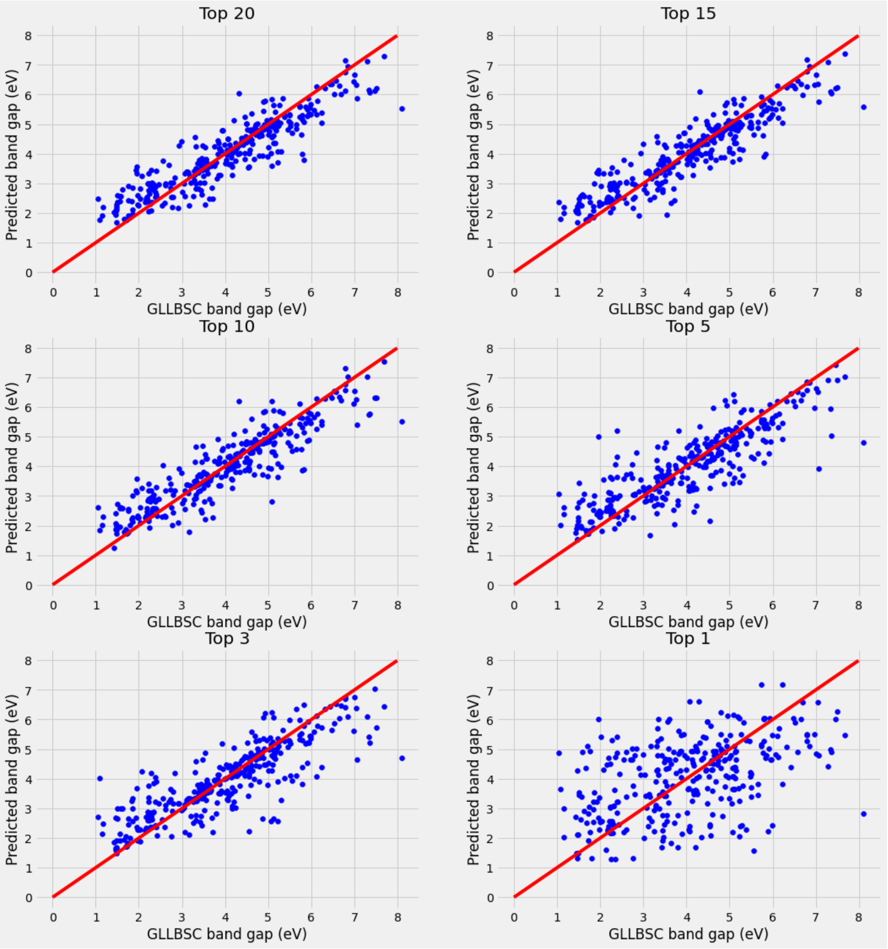

Figure 5. Parity plots of the predicted vs. GLLBSC-computed band gaps, obtained using different number of features with the highest importance scores. The parity line is shown in red.

Test and training set partition

Table 2 summarizes the model performance as a function of the different test-to-training set partition, ranging from 10/90 to 75/25. As expected, the test set accuracy decreases with the number of training set data points. The corresponding parity plots in **Fig. 6** also demonstrate larger extent of deviation from the parity line as the proportion of the training set decreases. Based on these results, we validate that the test/training ratio of 25/75 is sufficient in providing satisfactory accuracy (85.6%) and reasonable RMSE (0.64 eV).

Table 2. Model performance obtained with different test-to-training set partitions.

Test/training set ratio	10/90	20/80	25/75	40/60	50/50	75/25
Number of test set data points	131	262	327	523	653	980
Number of training set data points	1175	1044	979	783	653	326
Test set accuracy (%)	87.9	86.8	85.6	82.6	82.5	76.2
MAE (eV)	0.41	0.45	0.46	0.5	0.53	0.67

RMSE (eV)	0.57	0.63	0.64	0.7	0.74	0.88
NRMSE	0.07	0.08	0.08	0.08	0.09	0.11

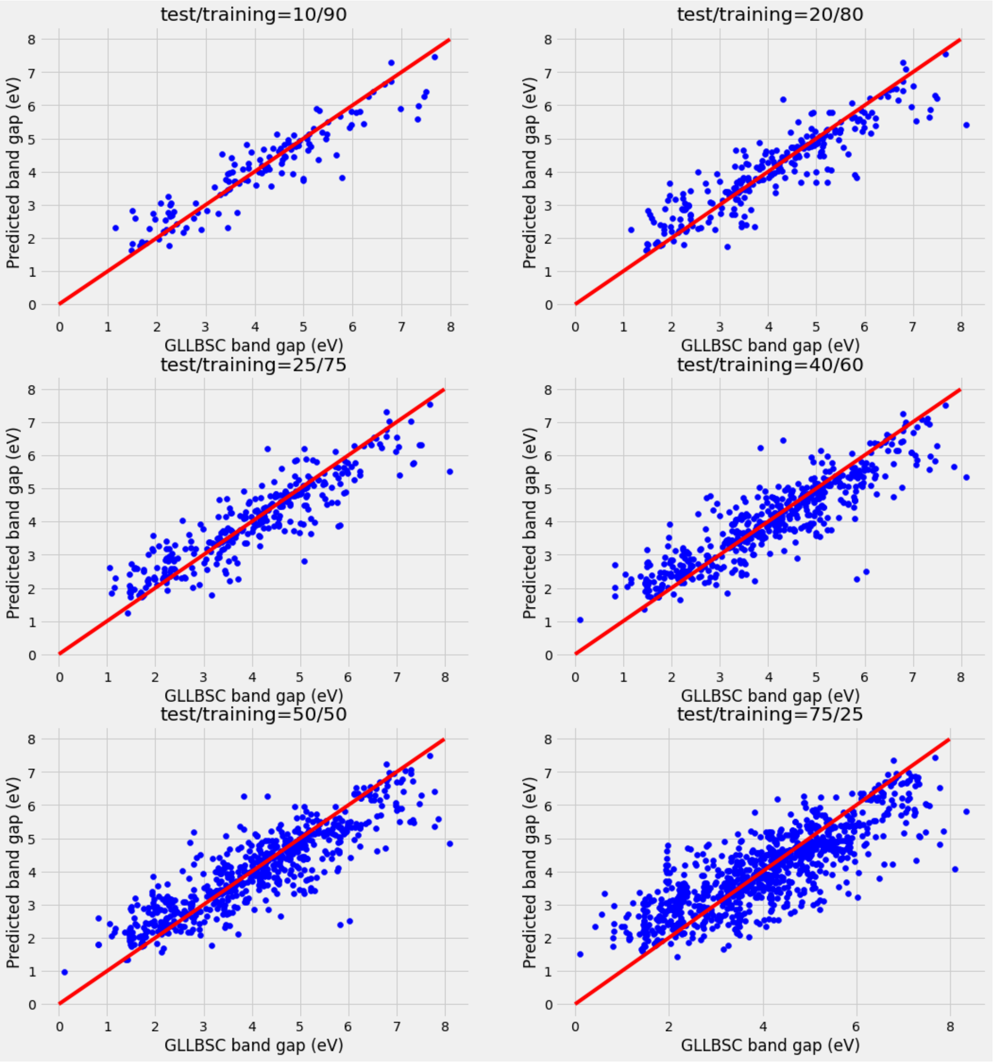

Figure 6. Parity plots of the predicted vs. GLLBSC-computed band gaps, obtained using different test-to-training set partitions. The parity line is shown in red.

Conclusions

Despite the widespread use of first-principles methods based on density functional theory (DFT) in materials science, it remains computationally costly and limited in its accuracy due to the approximation of the exchange-correlation functional. In this regard, machine learning presents a viable alternative for rapid prediction of materials electronic properties while retaining reasonable fidelity to DFT. This study has implemented random forest regression for prediction of the band gap of double perovskite compounds employing a dataset of 1306 GLLBSC-computed band gaps. Among the 20 physical descriptors, average bulk modulus, superconductivity temperature, and cation electronegativity exhibited the highest importance scores, which provide physically interpretable description in terms of the underlying electronic structure. Optimal model performance is obtained with the top 10 features and a test/training partition of 25/75, yielding a model accuracy of 85.6% and RMSE of 0.64 eV comparable to previous studies. Our results highlight the potential of machine learning regression for rapid and physically interpretable prediction of the electronic properties of functional materials.

Acknowledgments

This work was supported by Touch Education Technology Inc. We acknowledge scientific and editorial support from the Project Lead, J. S. Lim of Harvard University; technical support from the Project Support C. Zhang; and administrative support from C. Ding of Touch Education Technology Inc.

Author contributions

This work was led by J.Z. with support from Y.L. and X.Z. J.Z. performed machine learning, literature review, and drafted the manuscript. Y.L. performed parameter optimization, visualization, and literature review. X.Z. assisted with literature review and writing.

Competing financial interests

The authors declare no competing financial interests.

References

1. Saha-Dasgupta, T. Double perovskites with 3d and 4d/5d transition metals: Compounds with promises. *Materials Research Express* **7**, (2019) 014003.
2. Grabowska, E. Selected perovskite oxides: Characterization, preparation and photocatalytic properties-A review. *Applied Catalysis B: Environmental* **186**, 97–126 (2016).
3. Dey, A., Ye, J., De, A., Debroye, E., Ha, S., Bladt, E., Kshirsagar, A., Wang, Z., Yin, J., Wang, Y., Quan, L., Yan, F., Gao, M., Li, X., Shamsi, J., Debnath, T., Cao, M., Scheel, M., Kumar, S., Steele, J., Gerhard, M., Chouhan, L., Xu, K., Wu, X., Li, Y., Zhang, Y., Dutta, A., Han, C., Vincon, I., Rogach, A., Nag, A., Samanta, A., Korgel, B., Shih, C., Gamelin, D., Son, D., Zeng, H., Zhong, H., Sun, H., Demir, H., Scheblykin, I., Mora-Seró, I., Stolarczyk, J., Zhang, J., Feldmann, J., Hofkens, J., Luther, J., Pérez-Prieto, J., Li, L., Manna, L., Bodnarchuk, M., Kovalenko, M., Roeffaers, M., Pradhan, N., Mohammed, O., Bakr, O., Yang, P., Müller-Buschbaum, P., Kamat, P., Bao, Q., Zhang, Q., Krahne, R., Galian, R., Stranks, S., Bals, S., Biju, V., Tisdale, W., Yan, Y., Hoye, R., & Polavarapu, L. State of the Art and Prospects for Halide Perovskite Nanocrystals. *ACS Nano* **15**, 10775–10981 (2021).
4. Chen, P., Bai, Y., Wang, S., Lyu, M., Yun, J., Wang, L. In Situ Growth of 2D Perovskite Capping Layer for Stable and Efficient Perovskite Solar Cells. *Advanced Functional Materials* **28**, (2018) 1706923.

5. Wu, C., Zhang, Q., Liu, Y., Luo, W., Guo, X., Huang, Z., Ting, H., Sun, W., Zhong, X., Wei, S., Wang, S., Chen, Z., Xiao, L. The Dawn of Lead-Free Perovskite Solar Cell: Highly Stable Double Perovskite $Cs_2AgBiBr_6$ Film. *Advanced Science* **5**, (2018) 1700759.

6. Wang, H., Want, J., Pi, Y., Shao, Q., Tan, Y., & Huang, X. Double Perovskite $LaFe_xNi_{1-x}O_3$ Nanorods Enable Efficient Oxygen Evolution Electrocatalysis . *Angewandte Chemie* **131**, 2338–2342 (2019).

7. Du, K., Meng, W., Wang, X., Yan, Y. & Mitzi, D. B. Bandgap Engineering of Lead-Free Double Perovskite $Cs_2AgBiBr_6$ through Trivalent Metal Alloying . *Angewandte Chemie* **129**, 8270–8274 (2017).

8. Guo, M., Xu, X. & Xie, H. Predicting the band gap of binary compounds from machine-learning regression methods.

9. Sutherland, B. R. Solar Materials Find Their Band Gap. *Joule* **4**, 984–985 (2020).

10. Zhang, J., Yang, L., Qu, M., Qi, D., & Zhang, K. H. L.Wide Bandgap Oxide Semiconductors: from Materials Physics to Optoelectronic Devices. *Advanced Materials* **33** (2021) 2006230.

11. Bagayoko, D. Understanding density functional theory (DFT) and completing it in practice. *AIP Advances* **4**, (2014) 127104.

12. Gaa, I. E., Hohenbergt Ecole, P., Superzeure, I'aris, X., And, F. & Konnt, W. *PHYSICAL REVIEW* **136**, B868-B871 (1964).

13. Kuisma, M., Ojanen, J., Enkovaara, J. & Rantala, T. T. Kohn-Sham potential with discontinuity for band gap materials. *Physical Review B - Condensed Matter and Materials Physics* **82**, (2010) 115106.

14. Hai, X., Tahir-Kheli, J. & Goddard, W. A. Accurate band gaps for semiconductors from density functional theory. *Journal of Physical Chemistry Letters* **2**, 212–217 (2011).

15. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters* **77**, 3865-3868 (1996).

16. Lim, J. S., Saldana-Greco, D. & Rappe, A. M. Improved pseudopotential transferability for magnetic and electronic properties of binary manganese oxides from DFT+U+J calculations. *Physical Review B* **94**, (2016) 165151.

17. Dudarev, S. L., Botton, G. A., Savrasov, S. Y., Humphreys, C. J. & Sutton, A. P. Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDAU study. *Physical Review B* **57**, 1505-1509 (1998).

18. Franchini, C., Podloucky, R., Paier, J., Marsman, M. & Kresse, G. Ground-state properties of multivalent manganese oxides: Density functional and hybrid density functional calculations. *Physical Review B - Condensed Matter and Materials Physics* **75**, (2007) 195128.

19. van Setten, M. J., Weigend, F. & Evers, F. The GW-method for quantum chemistry applications: Theory and implementation. *Journal of Chemical Theory and Computation* **9**, 232–246 (2013).

20. Bagayoko, D. Understanding density functional theory (DFT) and completing it in practice. *AIP Advances* **4**, (2014) 127104.

21. Zhuo, Y., Mansouri Tehrani, A. & Brgoch, J. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *Journal of Physical Chemistry Letters* **9**, 1668–1673 (2018).

22. Rajan, A. C., Mishra, A., Satsangi, S., Vaish, R., Mizuseki, H., Lee, K., Singh, K. A., Machine-learning-assisted accurate band gap predictions of functionalized mxene. *Chemistry of Materials* **30**, 4031–4038 (2018).

23. Gritsenko, O., van Leeuwen, N., van Lenthe, E., & Baerends E. J. Self-consistent approximation to the Kohn-Sham exchange potential. *PHYSICAL REVIEW A* **51**, 1944-1954 (1995).
24. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* **5**, (2019) 83.
25. Guo, Z. & Lin, B. Machine learning stability and band gap of lead-free halide double perovskite materials for perovskite solar cells. *Solar Energy* **228**, 689–699 (2021).
26. Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. & Chica-Rivas, M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews* **71**, 804–818 (2015).
27. Biau, G. & Fr, G. B. Analysis of a Random Forests Model. *Journal of Machine Learning Research* **13**, 1063-1095 (2012).
28. Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., Chard, K., Asta, M., Persson, K., Snyder, G., Foster, I., Jain, A.Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **152**, 60–69 (2018).
29. Ong, S., Richards, W., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V., Persson, K., Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68**, 314–319 (2013).
30. Probst, P., Wright, M. N. & Boulesteix, A. L. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **9**, (2019) 1301.
31. Solé, X., Solé, S., Ramisa, A. & Torras, C. Evaluation of Random Forests on large-scale classification problems using a Bag-of-Visual-Words representation.
32. Segal, M. R. UCSF Recent Work Title Machine Learning Benchmarks and Random Forest Regression Publication Date Machine Learning Benchmarks and Random Forest Regression. (2003).
33. Chandrasekar, S. & Santhanam, S. A calculation of the bulk modulus of polycrystalline materials. *Journal of Materials Science* **24**, 4265-4267 (1989).
34. Li, K., Kang, C. & Xue, D. Electronegativity calculation of bulk modulus and band gap of ternary ZnO-based alloys. in *Materials Research Bulletin* **47**, 2902–2905 (2012).
35. Saha, S., di Cataldo, S., Amsler, M., von der Linden, W. & Boeri, L. High-temperature conventional superconductivity in the boron-carbon system: Material trends. *Physical Review B* **102**, (2020) 024519.
36. Panagopoulos, C. & Xiang, T. Relationship between the Superconducting Energy Gap and the Critical Temperature in High-T c Superconductors. *Physical Review Letters* **81**, 2336-2339 (1998).
37. Iczkowski, R. P. & Margrave, J. L. Electronegativity. *Journal of the American Chemistry Society* **83**, 3547-3551 (1961).
38. di Quarto, F., Sunseri, C., Piazza, S. & Romano, M. C. Semiempirical Correlation between Optical Band Gap Values of Oxides and the Difference of Electronegativity of the Elements. Its Importance for a Quantitative Use of Photocurrent Spectroscopy in Corrosion Studies. *J. Phys. Chem. B* **101**, 2519-2525 (1997).
39. Tran, F., Ehsan, S. & Blaha, P. Assessment of the GLLB-SC potential for solid-state properties and attempts for improvement. *Physical Review Materials* **2**, (2018) 023802.