Predicting the band gap of binary compounds from machine-learning regression methods

Mengbo Guo,[1] Xuyang Xu,[2] and Han Xie[3]

[1] Department of Physics and Technology, Wuhan University, Wuhan, Hubei Province, 430000, China
[2] Department of Physics, Peking University, Beijing, 100091, China
[3] Department of Physics, Nankai University, Tianjin, 300071, China

* Emails: 2019300001075@whu.edu.cn; 1700017717@pku.edu.cn; 1910335@mail.nankai.edu.cn
* The authors contributed equally to this work.

Abstract

Density functional theory (DFT) is a ubiquitous first-principles method, but the approximate nature of the exchange-correlation functional poses an inherent limitation for the accuracy of various computed properties. In this context, surrogate models based on machine learning have the potential to provide a more efficient and physically meaningful understanding of electronic properties, such as the band gap. Here, we construct a gradient boosting regression (GBR) model for prediction of the band gap of binary compounds from simple physical descriptors, using a dataset of over 4000 DFT-computed band gaps. Out of 27 features, electronegativity, periodic group, and highest occupied energy level exhibit the highest importance score, consistent with the underlying physics of the electronic structure. We obtain a model accuracy of 0.81 and root mean squared error of 0.26 eV using the top five features, achieving accuracy comparable to previously reported values but employing less number of features. Our work presents a rapid and interpretable prediction model for solid-state band gap with high fidelity to DFT and can be extended beyond binary materials considered in this study.

Introduction

According to the Schrödinger equation, a constrained particle occupies several eigenstates with quantized energy levels. For example, the ground state energy of the electron in a hydrogen atom is -13.6 eV, and the higher energy levels can be accessed by absorbing a photon whose wavelength is determined by the Rydberg equation. In solids with multiple atoms, the electronic states turn into continuous bands. The Fermi level divides the energy range into the valence and conduction bands. Insulators and semiconductors have bandgaps, the energy range where electrons are forbidden to occupy. To conduct electricity, electrons need to be excited across the bandgap from the valence band into the conduction band. Bandgaps, especially that of semiconductors, play a central role in many device applications, such as transistors,[1] light emitting diodes (LED),[2,3] photovoltaics,[4] and thermoelectric materials.[5] Therefore, accurate evaluation of bandgaps from computation is crucial to rational materials design.

Density functional theory (DFT) is widely used to calculate the electronic structure of solids. In the Kohn-Sham equation, the exact form of the exchange-correlation energy, corresponding to many-

body quantum interactions, remains unknown in terms of the electron density. As such, an approximation is needed, such as local density approximation (LDA) and generalized gradient approximation (GGA). Conventional DFT based on these approximations fail to calculate accurate bandgaps, and many insulators and semiconductors are predicted to be metallic.[6,7] As a result, more computationally demanding methods such as hybrid functionals[8] and GW method are needed.[9] A much cheaper and popular alternative is DFT+*U*, where Hubbard *U* on-site repulsion correction is applied to increase the bandgap. However, such corrections can lead to unphysical distortions of the electronic structure profile.[10] Recently, machine learning has seen an increase in applications to overcome many of these limitations of conventional DFT, e.g. machine-learning density functionals[11], as well as data-driven prediction of electronic properties such as band gaps at higher accuracy.[12] For example, regression methods such as support vector machine,[12] k-nearest neighbors,[13] and logistic regression,[14] have been used to successfully predict bandgaps using experimentally measured values.

In recent years, machine-learning assisted screening of functional materials has garnered much attention.[12,13,15] Machine learning methods have been used to predict a variety of materials properties, such as hardness, electrical conductivity,[16] melting temperatures,[17] phase stability,[18] and ionic conductivity.[19] The conventional trial-and-error method relies on the intuition and experience of researchers and requires a lot of time and resources. Although computational screening based on first-principles methods greatly facilitates the screening process, the computational cost remains very high. Machine learning is a viable alternative that can enable faster and more efficient exploration of the large materials space.

Here, we construct a gradient boosting regression model for prediction of the band gap of binary compounds from simple physical descriptors, using a dataset of more than 4000 DFT-computed band gaps. We build on a previous study of double perovskites that used random forest regression and linear ridge regression to predict DFT-computed band gaps with an improved accuracy compared to prior literature.[13] Out of 27 features, electronegativity, periodic group, and highest occupied energy level have the highest importance score, which provide a physically interpretable understanding of the band gap. Moreover, our model is capable of achieving accuracy comparable to previously reported values but employing less number of features.

Methods

Gradient boosting regression

In regression, the prediction function $\hat{F}$ is optimized by minimizing a loss function $L(y, F(x))$ for the true value y and the prediction $F(x)$:[20]

$$\hat{F} = \text{argmin}_F \mathbb{E}_{x,y}\big[L\big(y, F(x)\big)\big].$$

The gradient boosting method assumes a real-valued y and seeks an approximation $\hat{F}$ in the form of a weighted sum of functions $h(x)$, called base (or weak) learners, with weights $\gamma(x)$:

$$\hat{F} = \sum_{i=1}^{M} \gamma_i\,(x) h_i(x) + \text{const.}$$

The loss function is iteratively minimized using the steepest descent. The maximum-descent direction of the loss function F_m at iteration m is given by the function at the previous step $m-1$ subtracted by the loss function gradient:

$$F_m = \text{argmin}_\gamma \sum_{i=1}^{M} L(y_i, \gamma_i),$$

$$F_m = F_{m-1} + \text{argmin}_{h_m} \sum_{i=1}^{M} L(y_i, F_{m-1}(x_i) + h_m(x_i)),$$

$$\gamma_m = \text{argmin}_\gamma \sum_{i=1}^{M} L(y_i, F_{m-1}(x_i) - \gamma \nabla_{F_{m-1}} L\big(y_i, F_{m-1}(x_i)\big).$$

Dataset and features

We choose the dataset of DFT band gaps of 4096 binary compounds,[21] computed using the Perdew-Berke-Ernzerhof (PBE) parametrization[22] of the generalized gradient approximation (GGA) of the exchange-correlation functional. The dataset is available through the matminer package.[23] 27 atomic features are obtained using the Pymatgen (Python Materials Genomics) package,[24] including Pauling electronegativity, ionization potential, highest occupied atomic level, lowest unoccupied atomic level, and s-, p- and d-valence orbital radii of isolated neutral atoms.[25] Each element in a given compound is classified as either host or guest based on their relative composition from which weighted average of each feature is computed. All regression analysis is performed using the scikit-learn package.[26]

Results and discussion

Model selection and parameter optimization

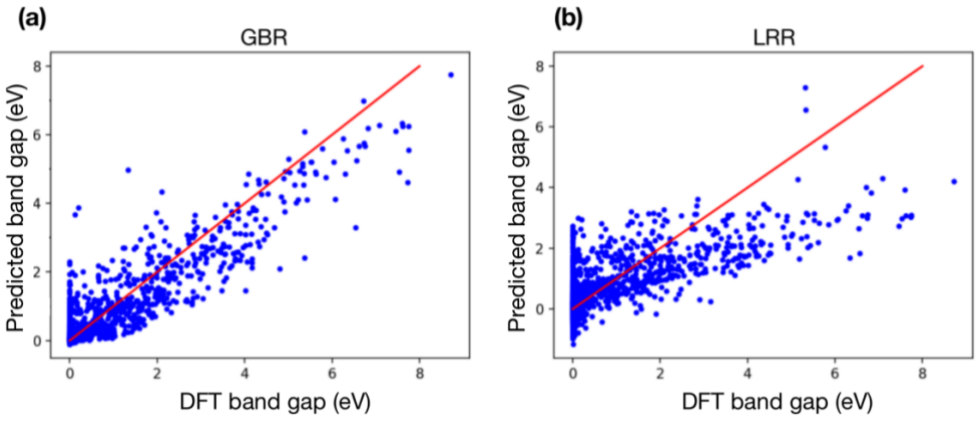

Figure 1. Predicted vs. DFT-computed band gaps on the training set using (**a**) gradient boosting regression (GBR; *n_estimators*=100, *learning_rate*=0.1, *subsample*=1) and (**b**) linear ridge regression (LRR; *alpha*=1). The parity line is shown in red. The data points lie closer to the parity line with GBR, indicating better performance.

We first compare linear vs. nonlinear regression. Linear ridge regression (LRR) and gradient boosting regression (GBR) models provide test set score of 0.50 and 0.85, respectively, as can be seen by data points lying closer to the parity line with GBR in **Fig. 1a**. This observation suggests that nonlinear regression is better suited to describe our dataset. GBR uses an ensemble of weak prediction models, typically decision trees, where boosting is used for optimization on a suitable cost function.[27] The main advantage of GBR is its resilience against overfitting. As such, we employ GBR to generate all results presented in the following discussions.

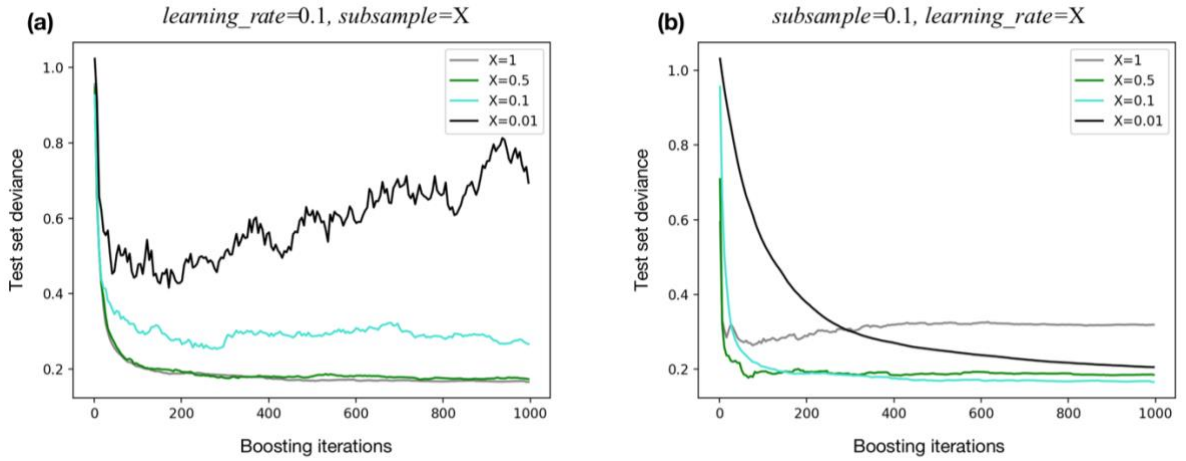

Figure 2. Test set deviance over 1000 boosting iterations using different set of parameters: (**a**) *subsample*=1 with varying *learning_rate* values; (**b**) *learning_rate*=0.1 with varying *subsample* values.

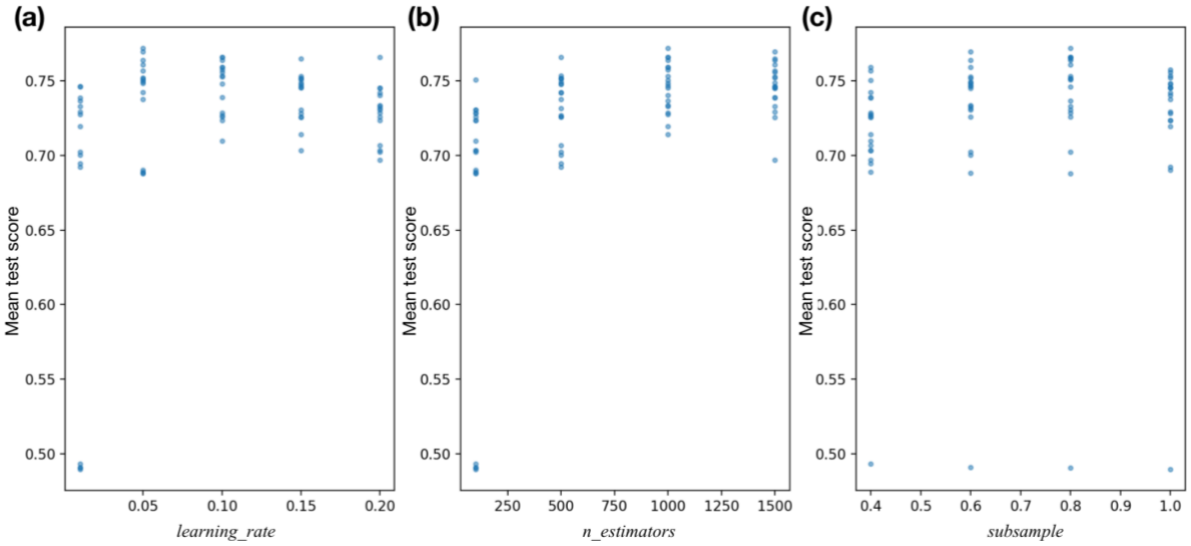

Figure 3. Cross validation and grid search are performed by varying one parameter at a time: (**a**) *learning_rate*, (**b**) *n_estimators*, and (**c**) *subsample*.

Next, we optimize three main parameters of the GBR model: (1) *learning_rate* for shrinking the contribution of each weak prediction trees; (2) *subsample* for the fraction of samples to be used for fitting the individual base learners; and (3) *n_estimators* for the number of boosting iterations. **Fig. 2** shows test set deviance while varying either *learning _rate* or *subsample* parameter at a time. Better model performance is seen with decreasing *learning_rate* and increasing *subsample*. Across both cases, the model remains fairly robust to overfitting and the deviance decreases exponentially, plateauing after

around 200th iteration. As such, better prediction can be obtained with larger values of *n_estimators*. From further grid search in **Fig. 3**, optimal parameters are chosen as *learning_rate*=0.1, *subsample*=1, and *n_estimators*=1000.

Training and test set partition

Table 1. Training and test set scores and test set root mean square error (RMSE) for models with different test/training ratios as considered in **Fig. 4**.

Test/training	10/90	20/80	25/75	40/60	50/50	75/25
Training set score	0.96	0.97	0.97	0.98	0.98	0.99
Test set score	0.86	0.84	0.84	0.81	0.78	0.73
RMSE (eV)	0.15	0.19	0.22	0.26	0.31	0.38

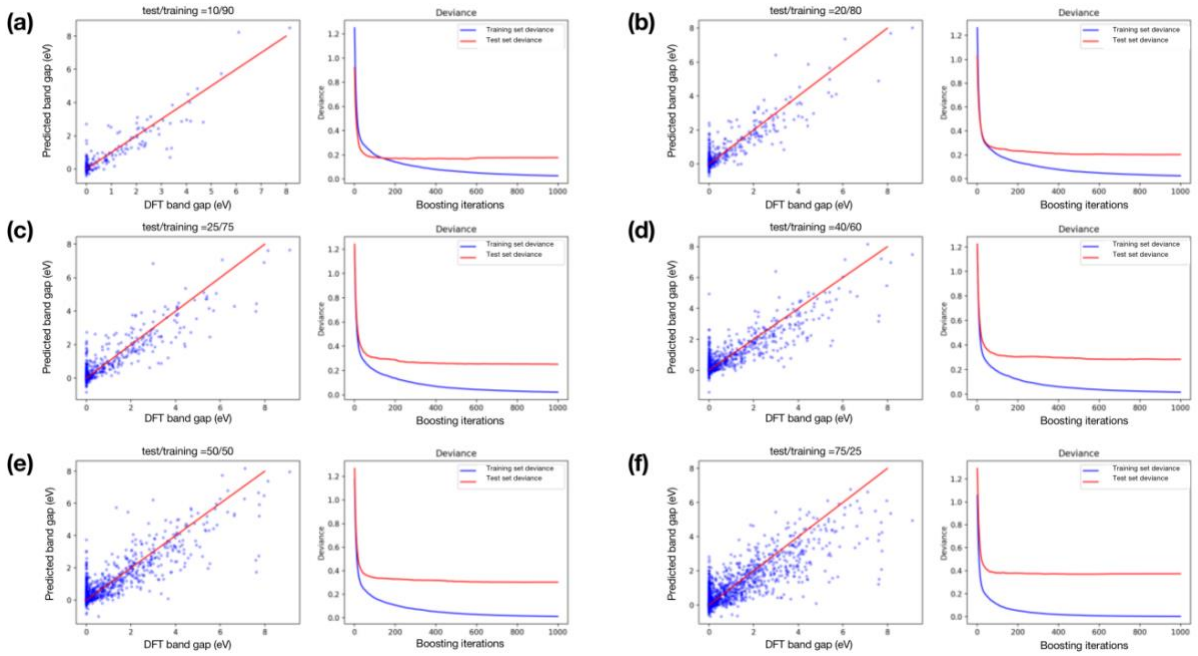

Figure 4. Predicted vs. DFT-computed band gap, with the parity line shown in red; and the training and test set deviance over 1000 boosting iterations; for models with different test/training ratios (**a-f**).

We investigate the effect of different training and test set partitions on the model performance, ranging from test/training ratio of 10/90 to 75/25. (**Fig. 4**). Increasing the relative proportion of the test set leads to a decrease in the test set score with concomitant increase of the root mean squared error (RMSE) and the training set score (**Table 1**). Based on this observation, we choose the test/training ratio of 20/80, which provides an optimal compromise of the training and test set score of 0.97 and 0.84, respectively, while maintaining a reasonable RMSE of 0.19 eV. Both test and training set deviances decrease across boosting iterations, the latter more so than the former, consistent with the observation of **Fig. 2**.

Feature importance and efficacy

Table 2. Training and test set scores and test set root mean square error (RMSE) for models with different number of features with the highest importance score as enumerated in **Fig. 5** and **6**.

Number of top features	27	17	11	5	3
Training set score	0.97	0.97	0.97	0.95	0.88
Test set score	0.84	0.80	0.78	0.81	0.66
RMSE (eV)	0.22	0.28	0.27	0.26	0.46

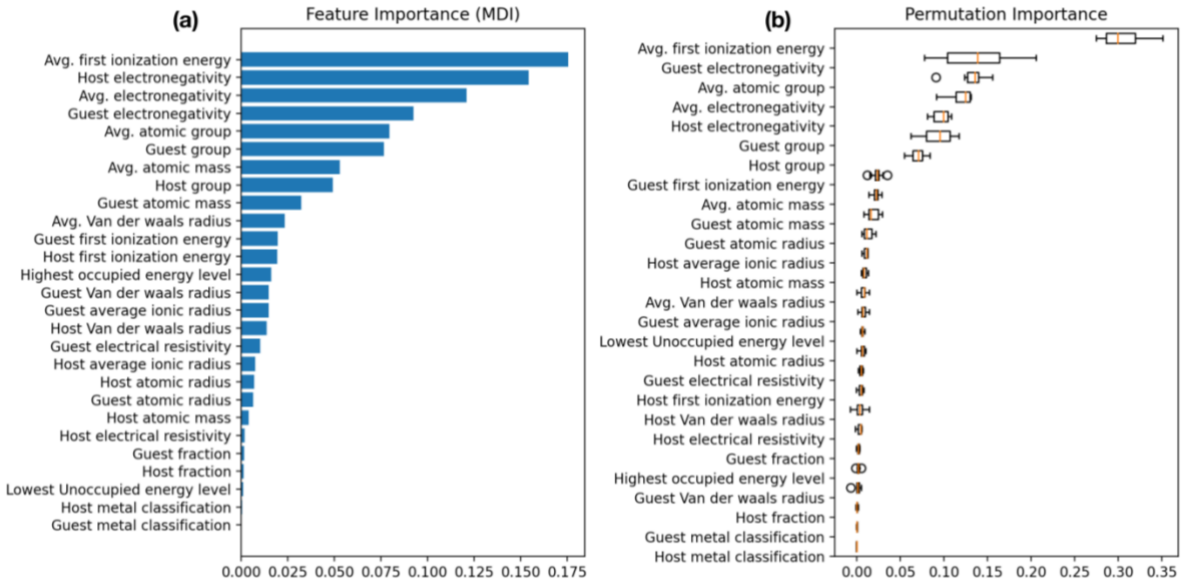

Figure 5. Feature importance scores for the 27 features considered in this study: (**a**) training set mean decrease in impurity (MDI) and (**b**) test set permutation importance.

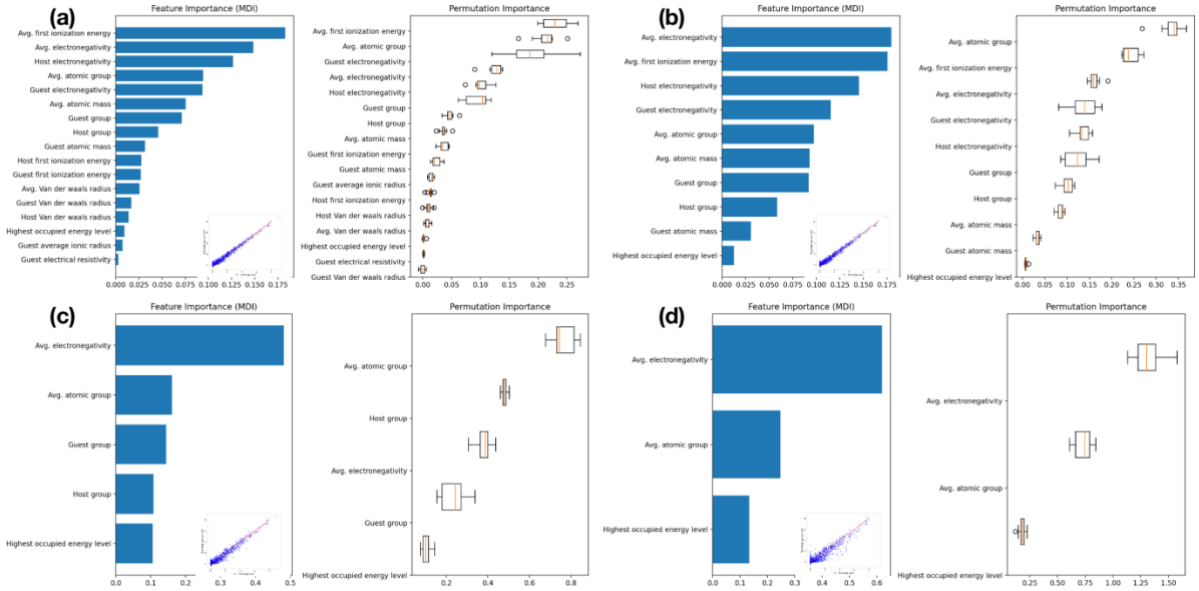

Figure 6. Training set mean decrease in impurity (MDI) and test set permutation importance for models with different number of features with the highest importance score as enumerated in **Fig. 5**.

We characterize the feature importance in terms of two metrics: (1) mean decrease in impurity (MDI) quantifies the mean and standard deviation of the accumulation of the impurity decrease within each tree, computed on the training set; (2) permutation importance quantifies the decrease in a model score when a single feature value is randomly shuffled, computed on the test set. Sequential reduction of the feature dimension is conducted to determine the optimal number of features with the highest importance score, ranging from 17 down to 3 features (**Fig. 5**). Overall, the model performance does not vary appreciably when the number of futures is reduced from 27 to 5 (**Table 2**), with the training and test set scores varying 0.95-0.97 and 0.78-0.84, respectively, and the RMSE varying 0.22-0.28 eV. The performance drops significantly when only three features are used, with the respective metrics of 0.88, 0.66, and 0.46 eV. Based on this observation, we conclude that 5 features are sufficient to retain high accuracy at levels comparable to previously reported model score of 0.919 using 12 features;[15] and cross-validation score of 96% and RMSE of 0.28 eV using 21 features.[27]

Three main types of features exhibited the highest importance score: electronegativity, periodic group, and highest occupied energy level. In terms of physics, the periodic potential of the constituent ions would influence the electronic structure and the overall conductivity as determined by the band gap. Electronegativity quantifies how strongly an atom attracts an electron within a chemical bond. In solid-state systems, electronegativity would correspond to the extent of electron localization around the constituent atoms, which influences the insulating character of the material. Such atomic feature has a close correlation with the periodic group, which is the column of the periodic table that an element belongs to. Transition metals toward the right side of the periodic table have higher number of protons, which increases the electronegativity, as well as higher number of electrons, which stabilizes the electrons and lowers the d-band center and the valence band maximum corresponding to the highest occupied energy level.

Conclusions

First-principles modeling based on density functional theory (DFT) is ubiquitous in computational materials science but remains limited in its accuracy and cost. Machine learning has the potential to provide efficient and physically interpretable model of crucial electronic structure properties, such as the band gap, while maintaining high fidelity to DFT. In this work, we construct a gradient boosting regression (GBR) model for prediction of the band gap of binary compounds from simple physical descriptors, using a dataset of over 4000 DFT-computed band gaps. Out of 27 features, electronegativity, periodic group, and highest occupied energy level exhibit the highest importance score, consistent with the underlying physics of the electronic structure. Using the top five features, we obtain a model accuracy of 0.81 and root mean squared error of 0.26 eV, which is comparable to previously reported values but employing fewer features. Our approach is broadly applicable for rapid and interpretable prediction of solid-state band gap for other types of materials beyond binary systems considered in this study.

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, L. Abraham of Panthéon-Sorbonne University; and administrative support from M. Clarke and X. Wang of Touch Education Technology Inc.

Author Contributions

H.X., X.X., and M.G. contributed equally to this work. H.X. performed coding and visualization. X.X. analyzed the regression algorithms. M.G. conducted literature review and writing. All authors contributed to manuscript preparation.

Competing Financial Interests

The authors declare no competing financial interests.

References:

[1] Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V. & Kis, A., Single-layer MoS2 transistors. *NAT NANOTECHNOL* **6** 147 (2011).

[2] Isayev, O. *et al.*, Universal fragment descriptors for predicting properties of inorganic crystals. *NAT COMMUN* **8** 15679 (2017).

[3] Curtarolo, S. *et al.*, AFLOW: An automatic framework for high-throughput materials discovery. *COMP MATER SCI* **58** 218 (2012).

[4] Polman, A., Knight, M., Garnett, E. C., Ehrler, B. & Sinke, W. C., Photovoltaic materials: Present efficiencies and future challenges. *SCIENCE* **352** d4424 (2016).

[5] XU, Y. *et al.*, New materials band gap prediction based on the high-throughput calculation and the machine learning. *SCIENTIA SINICA Technologica* **49** 44 (2018).

[6] Seidl, A., Gorling, A., Vogl, P., Majewski, J. A. & Levy, M., Generalized Kohn-Sham schemes and the band-gap problem. *Phys Rev B Condens Matter* **53** 3764 (1996).

[7] Perdew, J. P., Density functional theory and the band gap problem. *INT J QUANTUM CHEM* **28** 497 (1985).

[8] Garza, A. J. & Scuseria, G. E., Predicting Band Gaps with Hybrid Density Functionals. *The Journal of Physical Chemistry Letters* **7** 4165 (2016).

[9] Gerosa, M. *et al.*, Electronic structure and phase stability of oxide semiconductors: Performance of dielectric-dependent hybrid functional DFT, benchmarked against GW band structure calculations and experiments. *PHYS REV B* **91** 155201 (2015).

[10] Lim, J. S., Saldana-Greco, D. & Rappe, A. M., Improved pseudopotential transferability for magnetic and electronic properties of binary manganese oxides fromDFT+U+J calculations. *PHYS REV B* **94** 165151 (2016).

[11] Li, L. *et al.*, Understanding machine-learned density functionals. *INT J QUANTUM CHEM* **116** 819 (2016).

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

[13] Pilania, G. *et al.*, Machine learning bandgaps of double perovskites. *SCI REP-UK* **6** 19375 (2016).

[14] Li, Y. *et al.*, Bandgap tuning strategy by cations and halide ions of lead halide perovskites learned from machine learning. *RSC ADV* **11** 15688 (2021).

[15] Yang, X., Li, L., Tao, Q., Lu, W. & Li, M., Rapid discovery of narrow bandgap oxide double perovskites using machine learning. *COMP MATER SCI* **196** 110528 (2021).

[16] Zhao, Q. *et al.*, Machine learning-assisted discovery of strong and conductive Cu alloys: Data mining from discarded experiments and physical features. *MATER DESIGN* **197** 109248 (2021).

[17] Seko, A., Maekawa, T., Tsuda, K. & Tanaka, I., Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids. *PHYS REV B* **89** 54303 (2014).

[18] Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M., Big Data of Materials Science: Critical Role of the Descriptor. *PHYS REV LETT* **114** 105503 (2015).

[19] Sendek, A. D. *et al.*, Holistic computational structure screening of more than 12 000 candidates for solid lithium-ion conductor materials. *ENERG ENVIRON SCI* **10** 306 (2017).

[20] Friedman, J. H., Greedy Function Approximation: A Gradient Boosting Machine. *ANN STAT* **29** 1189 (2001).
[21] Jain, A. *et al.*, {The Materials Project: A materials genome approach to accelerating materials innovation}. *APL MATER* **1** 11002 (2013).
[22] Burke, K., Ernzerhof, M. & Perdew, J. P., Generalized Gradient Approximation Made Simple. *PHYS REV LETT* **77** 3865 (1996).
[23] Ward, L. *et al.*, Matminer: An open source toolkit for materials data mining. *COMP MATER SCI* **152** 60 (2018).
[24] Ong, S. P. *et al.*, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *COMP MATER SCI* **68** 314 (2013).
[25] Friedman, J. H., Stochastic gradient boosting. *COMPUT STAT DATA AN* **38** 367 (2002).
[26] Pedregosa, F. *et al.*, Scikit-learn: Machine Learning in {P}ython. *J MACH LEARN RES* **12** 2825 (2011).
[27] Friedman, J. H., Stochastic gradient boosting. *COMPUT STAT DATA AN* **38** 367 (2002).
[28] Wang, T., Tan, X., Wei, Y. & Jin, H., Accurate bandgap predictions of solids assisted by machine learning. *Materials Today Communications* **29** 102932 (2021).