Applied Physics Letters ARTICLE pubs.aip.org/aip/apl

Accurate prediction of dielectric properties and bandgaps in materials with a machine learning approach

Cite as: Appl. Phys. Lett. **125**, 152905 (2024); doi: 10.1063/5.0223890
Submitted: 18 June 2024 · Accepted: 29 September 2024 ·
Published Online: 10 October 2024

View Online Export Citation CrossMark

Yilin Hu,[1] Maokun Wu,[1,a)] Miaojia Yuan,[1] Yichen Wen,[1] Pengpeng Ren,[1] Sheng Ye,[1] Fayong Liu,[2] Bo Zhou,[3] Hui Fang,[4] Runsheng Wang,[5] Zhigang Ji,[1,a)] and Ru Huang[5]

AFFILIATIONS

[1]National Key Laboratory of Science and Technology on Micro/Nano Fabrication, Shanghai Jiao Tong University, Shanghai 200240, China
[2]College of Electronic Engineering, Ocean University of China, Qingdao 266404, China
[3]School of Computer Science and Mathematics, Liverpool John Moores University, Byrom Street Liverpool L3 3AF, UK
[4]Department of Computer Science, Loughborough University, Loughborough LE11 3TU, UK
[5]School of Integrated Circuits, Peking University, Beijing 100871, China

[a)]**Authors to whom correspondence should be addressed:** maokunwu@sjtu.edu.cn and zhigangji@sjtu.edu.cn

ABSTRACT

The conventional approach to exploring suitable dielectrics for future logic and memory devices relies on first-principle calculations, which are expensive and time-consuming. In this work, we adopt a data-driven machine learning (ML)-based approach to build a model for predicting these properties. By incorporating structural information into the input descriptors, we achieve record-high accuracy in predicting the dielectric constant, with the coefficients of determination (R^2) of 0.886 and root mean square error (RMSE) of 0.083. Additionally, we achieve high predictions for the bandgap, with accuracies of 0.832 and 0.533 for R^2 and RMSE, respectively. The features corresponding to specific properties are analyzed to obtain physical insights. Finally, we employ first-principle calculations to validate the feasibility of this model. This work proposes a highly efficient approach for using ML to predict material properties.

The DRAM capacitor, which consists of a metal–insulator–metal (MIM) structure, must store sufficient charge for sufficient time to maintain a reasonable refresh rate. Following the continuous and rapid scaling of electronic devices to the nanometer scale, silicon dioxide (SiO_2, $\kappa = 4$), as an archetypical dielectric material, exhibits a dominant leakage current due to the quantum tunneling effect.[1] In contrast, high-κ dielectrics are expected, since they allow for greater thickness when maintaining the same capacitance, thereby suppressing the leakage current.[2] For example, HfO_2 is widely used as the gate dielectric in DRAM owing to its high-κ, with ~25 and the compatibility with complementary-metal-oxide-semiconductor (CMOS) technology.[3] Even so, a large bandgap (E_g) is also required to suppress the charge injection from the electrode into the dielectric.[4] However, κ and E_g present an inversely proportional relationship within a limited material dataset, as reported.[5] Currently, dielectric materials have been continuously discovered with the development of technology. Therefore, exploring the dielectric materials with high-κ values and wide Eg simultaneously is critical to meet the requirement of low leakage current for next-generation electronic devices such as memory devices and capacitor-based energy storage.

Remarkably, for the discovery of materials, a time-consuming process is required in the experimental aspect. In contrast, first-principles calculations have been considered as a powerful discovery tool for exploring material properties. Utilizing a high-throughput method with first-principles calculations, Yim *et al.* investigated the dielectric constants and E_g of more than 1800 binary and ternary oxides.[4] Umeda *et al.* computed the dielectric constants of 2393 oxides, suggesting 24 compounds with dielectric constants larger than 100.[6] However, powerful computational resources are required in high-throughput screening process since the calculation of κ can be computationally intensive. In contrast, data-driven machine learning, as an emerging discipline, has attracted widespread attention, since it can establish a simple mapping between fundamental physical parameters and target results, accelerating the discovery and design of materials in

a relatively short period of time. Morita *et al.* achieved the coefficients of determination (R^2) of 0.86 in predicting dielectric constants through constructing support vector regression (SVR) models. However, the root mean square error (RMSE) is as high as 0.99, reflecting a large deviation between predicted and actual values.[7] Recently, the prediction accuracy of κ can be improved by separating κ into electronic (ε_{el}) and ionic contributions (ε_{ion}) in machine learning models, respectively.[8] This is due to the fact that the accuracy of dominating ε_{ion} in κ is relatively low, resulting in a big error in the overall κ.[8] Kim *et al.* achieved an accuracy of 0.68 in predicting the ε_{ion} using a gradient boosting regression (GBR) model for perovskite-type oxides (ABO_3).[9] Takahashi *et al.* constructed a regression model to predict the ε_{ion} for ground-state oxides, achieving an improved accuracy of 0.73,[8] which is still not satisfactory. Further improvement to minimize error is required. Therefore, exploring improved schemes for predicting κ utilizing ML is critical to accelerate the identification of dielectric oxides.

Based on the above-mentioned motivations, we employed the support vector regression (SVR) model to predict the dielectric constants and E_g of binary and ternary oxides regarding its simple synthesis process. Compared with deep learning models, our current dataset only contains 722 materials, which is not enough to train a reliable deep-learning model. Therefore, classical machine learning models such as SVR are more suitable for small datasets.[10–14] In this work, the structural information is considered thoroughly owing to its significant role in determining ionic contribution. It is found that the prediction accuracy can be improved when the structural information is used as input descriptors, especially for the ionic contribution. Furthermore, the significance of the descriptors has been analyzed to gain physical insights into the dielectric constants. Additionally, several materials have been evaluated to validate the feasibility of this model by comparing it with DFT calculations. This work provides a useful reference in predicting dielectric constants and promoting the discovery of dielectrics.

As depicted in Fig. 1, the orange arrows indicate the expected area with large R^2 and small RMSE. R^2 indicates the relationship strength between the dependent variable and regression models on a 0–1 scale.[15,16] RMSE reflects the deviation between predicted and actual values, and smaller RMSE indicates better performance.[17–19] Figure 1(a) presents the comparison results of R^2 and RMSE between our model and the data from the literature. It is found that this model in our work can achieve better prediction results with R^2 (0.886) and RMSE (0.083), which is superior to the reported data.[7,20–22] This can be ascribed to the improved accuracy of the dominated ionic contribution, as shown in Fig. 1(a). The ε_{ion} achieves the highest R^2 score (0.798) compared to previous data, and the RMSE (0.163) of ε_{ion} is also comparable to the current accuracy.[8,9] In addition, as depicted in Fig. 1(b), we conducted a comparative analysis of the computational time-cost between the ML model and DFT calculations. Take $YGaO_3$ as an example; the calculation of the Born effective charge and phonon frequencies for dielectric constants can be computationally intensive when determining κ. In contrast, a significant reduction in computational time is achieved by employing this model in our work. Additionally, for bandgap estimation, utilizing ML for E_g prediction is significantly better than DFT calculations. This clearly illustrates that employing machine learning for predicting material properties can lead to substantial savings in both computational time and resources.

In this work, 722 binary and ternary oxides with dielectrics constants and E_g values are obtained as a database from Materials Project (MP) and relevant reports.[8,23] The support vector regression model is built based on statistical theory.[24–26] It exhibits strong generalization capabilities and captures robust relationships between features in small-sample scenarios. We compared various kernel functions, including linear, polynomial, radial basis function (RBF), and sigmoid kernels. After careful consideration, we selected the RBF kernel due to its superior performance in fitting the data and minimizing errors, which is realized using the Scikit-learn software package.[27,28] Furthermore, we systematically used grid search to explore a wide range of hyperparameter values. The final hyperparameters were set to C = (100, 100, 10) and gamma = (0.01, 0.01, 0.1) for electronic contribution, ionic contribution, and bandgap, respectively. The detailed heatmaps of different hyperparameters C and gamma in the grid search are given in the supplementary material.

To construct the machine learning model, the composition and structural information of these materials have been extracted to construct 126 characteristic features using the MATMINER[29] and PYMATGEN[30] packages. Remarkably, due to the pronounced sensitivity of phonon frequencies to ε_{ion} and structural parameters such as bond lengths, these factors play a significant role in determining phonon behavior in materials.[31–33] Meanwhile, since the ionic contribution plays a dominant role in the overall dielectric constant, enhancing

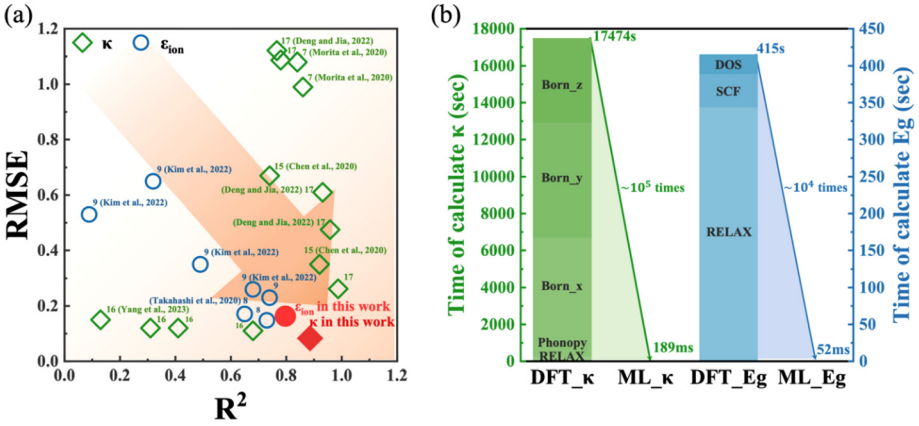

FIG. 1. (a) The comparison graph between our model and existing models in terms of R^2 score and root mean square error (RMSE). The rhombus represents the predicted values of the dielectric constant. Compared to other models, our model ($R^2 = 0.887$) has reached the optimal level. The circles represent the ε_{ion}. (b) The comparison graph between the computational times of density functional theory (DFT) calculations and machine learning (ML) predictions. As an example with 10 atoms of $YGaO_3$, predicting E_g saved $\sim 10^5$ times, while predicting dielectric constants saved $\sim 10^4$ times.

its accuracy is crucial for improving the overall precision of the dielectric constant. Therefore, we extracted a variety of structural features from the material's structural files as input parameters such as lattice constants, bond lengths, bond angles, and space groups. To train the model, we allocate 90% for the training set and 10% for the testing set. To improve the accuracy of the ML model, grid search technology is adopted to select the optimized hyperparameters.[34] Similarly, the R^2 score and RMSE are considered as performance metrics to evaluate the reliability of the ML model [Fig. 2(a) shows the workflow for ML]. The selection of an appropriate feature dimension is crucial for enhancing the precision of machine learning, which can capture material properties. Meanwhile, the number of the selected features should be fewer than the dataset itself to avoid the curse of dimensionality.[35,36] Thus, feature correlation analysis is initially conducted through calculating the Pearson correlation coefficient between each pair of features. When two features have an absolute Pearson correlation coefficient (ρ) greater than 0.9, one of the features is removed from the dataset. The features are then reduced to 82 [shown in Fig. 2(b)]. Subsequently, utilizing the embedded random forest regression algorithm, we construct the feature selection library efficiently and rapidly in the Scikit-learn software.[28,37] After multiple iterations, the final features for ε_{el}, ε_{ion}, and E_g are determined to be 15, 20, and 10, respectively.

To accurately quantify feature importance, we selected random forest regression to evaluate the impact of each feature on the model's results due to its distinguished advantages for data mining, handling high dimensional data, and detecting feature significance. Meanwhile, to ensure the accuracy of the evaluation results, we performed 500 iterations and averaged the importance scores for each feature.[37] Herein, the top five features selected are illustrated in Figs. 3(a)–3(c) according to their scores. For ε_{el}, the highest occupied molecular orbital (HOMO) energy is found to be the most important descriptor [Fig. 3(a)]. The energy difference between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) and the E_g show some degree of positive correlation in the statistical results.[8,38] The inverse relationship between E_g and ε_{el} has been proved in many studies.[5,7,39] The main reason is due to the calculation of the electronic static node tensor, where the three primary influencing factors are the joint density of states (DOS) composed of valence band and conduction band DOS, the extent of valence band to conduction band transition under the influence of an electric field, and the E_g and width of the valence and conduction bands.[8,40,41] Due to the dominated role of ε_{ion} in the dielectric constant and the strong dependence of the ionic static dielectric tensor on lattice parameters and internal atomic coordinates, feature importance ranking reveals the crucial significance of bond length features in determining ε_{ion} [Fig. 3(b)]. In addition, for bandgap prediction, it is evident that d valence electrons play a vital role [Fig. 3(c)]. According to the linear combination theory of atomic orbitals, the energy bands are determined by the energy difference between the bonding and antibonding orbitals. Thus, the E_g is closely related to the d orbitals of the elements. The quantity of d valence electrons can significantly elevate the conduction band energy level through s–d and p–d coupling.[42] Figure 3(d) shows the relationship between the E_g and the number of d valence electrons, confirming that the E_g decreases with an increase in the number of d valence electrons. This observation demonstrates the consistency between the model and theoretical predictions.

The predicted results through ML training are depicted in Fig. 4. Herein, the RBF kernel function is selected due to the higher accuracy. It can be observed that the predicted ε_{el} from the well-trained model are uniformly distributed around the y = x function, aligning with the data obtained from the input dataset. For ε_{el}, the R^2 of the test set is 0.935, and the RMSE is 0.039. Similarly, for ε_{ion}, the R^2 is 0.789 and the RMSE is 0.132. It is evident that the ε_{ion} model has achieved optimal performance compared to previous research.[8] Figure 4(c) presents a comparison between the model and DFT across the entire k, with an R^2 of 0.886 and an RMSE of 0.083, demonstrating good consistency that surpasses current literature reports. This improvement can be attributed to the increased accuracy of the ionic contributions through considering structural information. This indicates the reliability and accuracy of this model in our work. In addition, the accuracies of the bandgap predictions are 0.832 and 0.533 for R^2 and RMSE [shown in Fig. 4(d)], which is also comparable with the reported data.[43,44]

In order to demonstrate the feasibility of this model proposed in this work, first-principles calculations are employed using the PWmat software in the plane wave pseudo-potential formalism to calculate the κ and gap.[45] The polarization response entirely determines the dielectric constant. The linear-response method based on the density functional perturbation theory (DFPT) is used to obtain phonon modes and Born effective charges.[46] In this context, the dielectric constant is the sum of contributions from both electrons and ions. Due to the ionic part of κ being sensitive to the low-frequency phonon modes,[4] dielectric constant calculations employ the norm-conserving pseudopotential of local density approximation (LDA+SG15) to reduce the errors. A structural optimization is stopped until all residual force on each atom is smaller than 10^{-3} eV/Å.

The dielectric constant is given by

$$\varepsilon_{\alpha\beta}(\omega) = \varepsilon^{\infty} + \frac{4\pi}{\Omega}\sum_{n}\frac{f_{\alpha}^{*n} * f_{\beta}^{n}}{\omega_n^2 - \omega^2 - i\omega\Gamma}, \quad (1)$$

where ε^{∞} represents the ε_{el}, and the remainder represents ε_{ion}. Ω, ω, and f_{α}^{*n} represent the oscillator strength, the unit-cell volume, and the frequency of the infrared-active phonon, respectively. The detailed calculation steps are shown in Fig. 5(a).

To estimate the feasibility of this model, $XZrO_3$ (X = Mg, Ga, Sr, Ba) as ternary oxides and X_2O_3 (X = Sc, Y, La) as binary oxides are chosen from the candidate materials for DFT validation, and these materials are not included in the dataset. The crystal structures are illustrated in Fig. 5(b). The projected density of states is displayed for $XZrO_3$ (X = Mg, Ga, Sr, Ba) and X_2O_3 (X = Sc, Y, La), as shown in Fig. 5(c). For ternary oxides, the bandgap values are 4.11, 4.12, 3.74, and 3.28 eV, respectively. From Fig. 5(c), it can be observed that the valence band maximum is mainly contributed by O-2p orbitals, and the conduction band maximum is mainly contributed by Zr-4d orbitals. For binary oxides, the bandgap values are 3.91, 4.22, and 3.63 eV, respectively. From Fig. 1(b), it can be observed that the valence band maximum is mainly contributed by O-2p orbitals, and the conduction band maximum is mainly contributed by Sc-3d, Y-4d, and La-5d orbitals, respectively. In Fig. 5(d), we have also plotted the comparative relationship between E_g and κ utilizing this model in our work and DFT calculations. It is found that the consistency is well depicted in Fig. 5(d) for binary and ternary candidates, further demonstrating the feasibility of this model in predicting κ and Eg in our work.

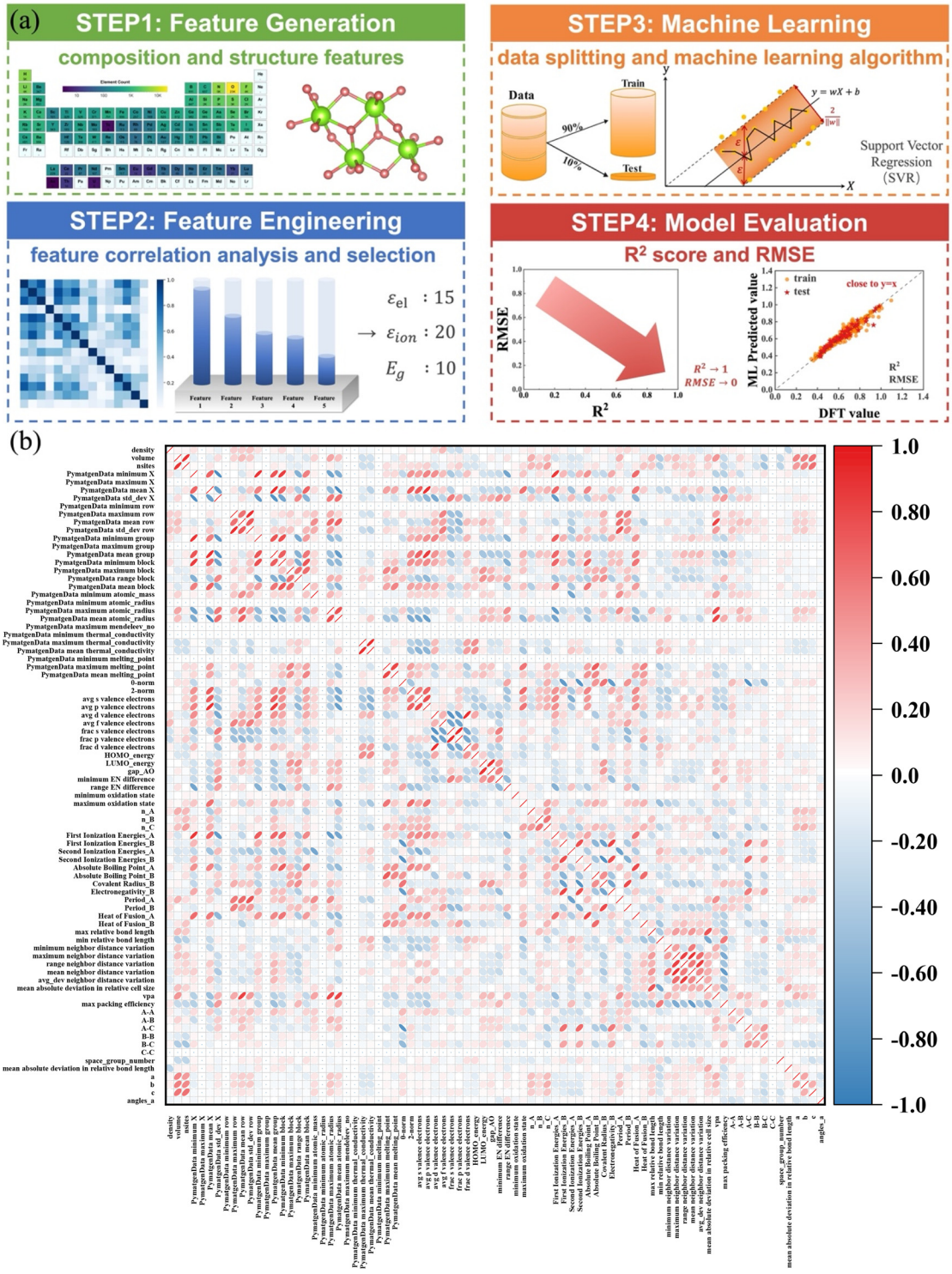

FIG. 2. (a) Workflow for constructing a machine learning (ML) model. Starting from the initial composition and structure of materials, through feature engineering and ML training, we ultimately evaluated using R^2 and RMSE. (b) The selected 82 feature descriptors of the oxides. The deeper the red, the stronger the positive correlation; the deeper the blue, the stronger the negative correlation. The direction of the ellipse indicates the sign of the correlation (positive or negative), and the flatness of the ellipse represents the strength of the correlation. The more circular the ellipse, the weaker the correlation.

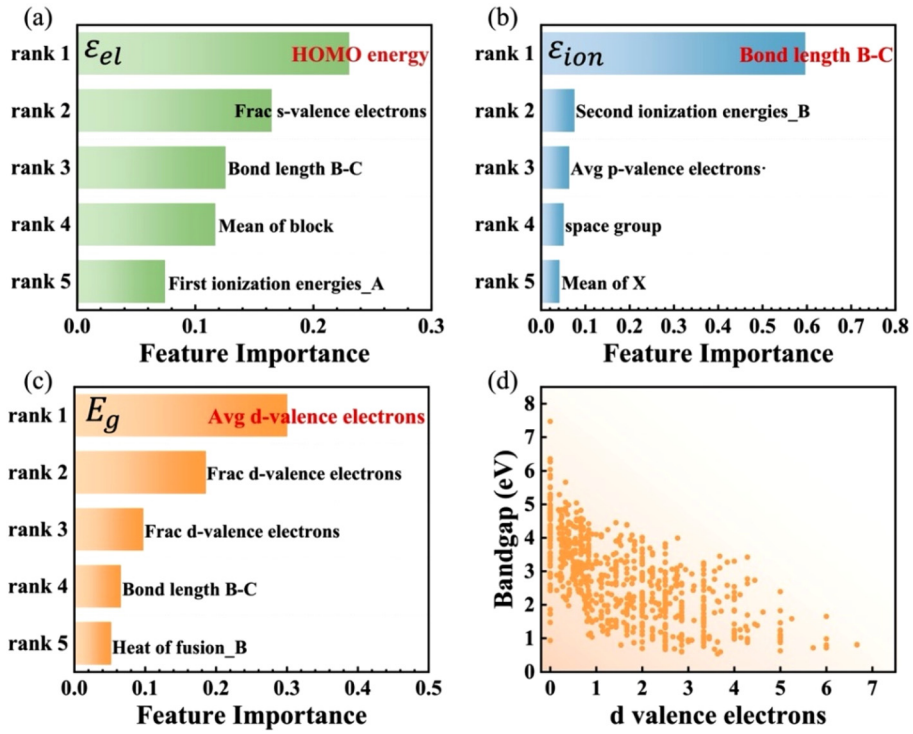

FIG. 3. The relative importance of the first five features screened for the (a) electronic contributions (ε_{el}), (b) ionic contributions (ε_{ion}), and (c) bandgap (E_g). One of the influential features affecting ε_{ion} is the B-C feature, which represents the bond length between the B atom and the O atom in ternary oxides ($A_xB_yO_z$). (d) Relationship between the d valence electrons and the bandgap predicted by machine learning, confirming that the E_g decreases with an increase in the number of d valence electrons.

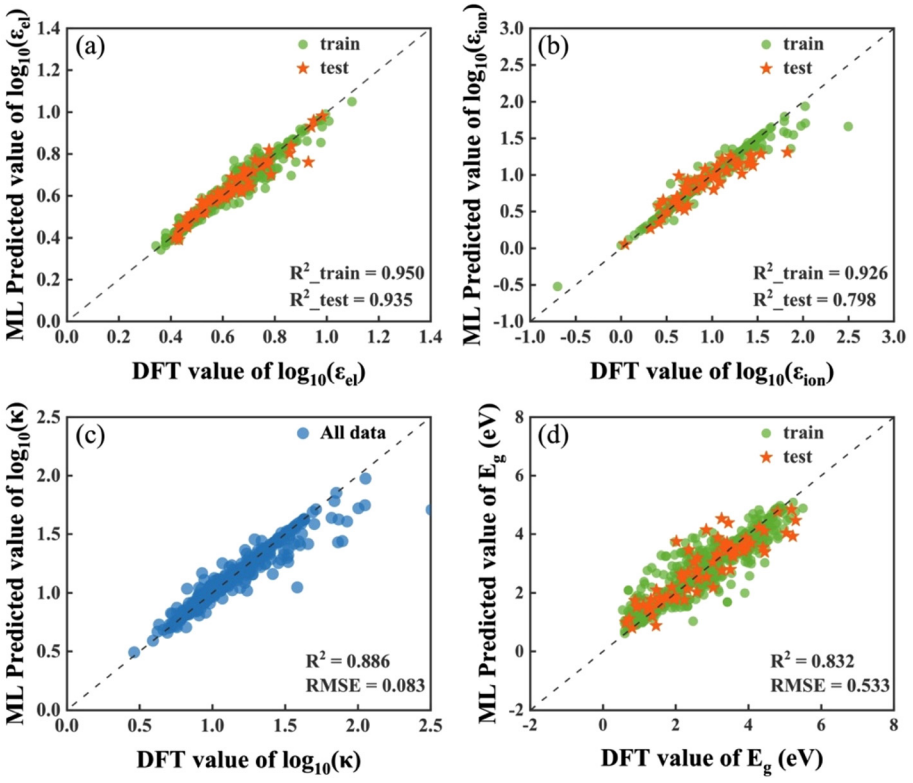

FIG. 4. Scatter plots of the (a) electronic dielectric constants (ε_{el}), (b) ionic dielectric constants (ε_{ion}), (c) κ, and (d) E_g between the reference values and ML predicted values, respectively. The orange stars and green crosses represent the training data (used to construct the model) and test data (not used to construct the model).

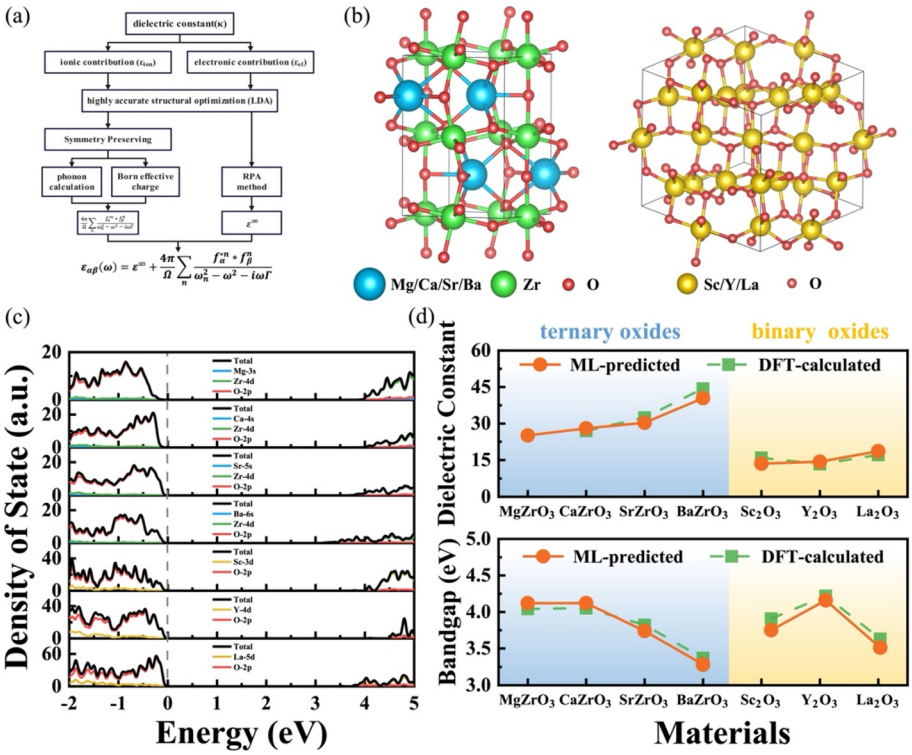

FIG. 5. (a) Workflow for calculating κ using PWmat. (b) Crystal structures of the ternary oxide $XZrO_3$ (X = Mg, Ga, Sr, and Ba) and the binary oxide X_2O_3 (X = Sc, Y, La). (c) The density of states (DOS) and partial DOS (PDOS) of $XZrO_3$ and X_2O_3. The Fermi level is set to be 0.0 eV. (d) Comparison of DFT-calculated values and ML-predicted values for the selected structures. The orange color indicates the predicted results using this model, while the green color is the results of the DFT calculations.

Remarkably, to further improve the accuracy of our proposed model, we can focus on providing more training samples and improve the model's generalization capability. For example, by conducting additional high-throughput DFT calculations to provide more data points for model training, it is expected to improve its ability to generalize.[11,47,48] In addition, data augmentation can be considered since it can effectively increase the size of the dataset by creating transformed versions of the original data.[49–51]

To summarize, by integrating ML with first-principles calculations, we have established models for the κ and E_g of binary and ternary oxides. Through a comparative analysis of computational time, our ML-based predictions of oxide properties exhibit a time saving $\sim 10^5$ times and 10^4 times for k and Eg calculations. Furthermore, through considering structural information as descriptors in the machine learning model, R^2 (0.886) and RMSE (0.083) for predicting k achieve a superior level owing to the sensitivity of phonon frequencies on structural parameters. For validation purposes, $XZrO_3$ (X = Mg, Ga, Sr, Ba) and X_2O_3 (X = Sc, Y, La) as candidate materials have been demonstrated. The results obtained from DFT calculations align closely with the predictions made using the ML model. This research aids in rapid material screening, significantly shortening the development cycle of next-generation microelectronic devices.

See the supplementary material for the detailed heatmaps of different hyperparameters C and gamma in the grid search for the electronic contribution, ionic contribution, and bandgap, respectively.

This work was supported by the National Natural Science Foundation of China (Nos. 62304136, 62027818, 61874034, and 11974320).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Yilin Hu: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Methodology (equal); Writing – original draft (equal); Writing – review & editing (equal). **Maokun Wu:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Methodology (equal); Supervision (supporting); Writing – review & editing (supporting). **Miaojia Yuan:** Conceptualization (supporting); Data curation (supporting); Formal analysis (supporting); Methodology (supporting). **Yichen Wen:** Data curation (supporting); Methodology (supporting). **Pengpeng Ren:** Conceptualization (supporting); Data curation (supporting); Formal analysis (supporting); Methodology (supporting). **Sheng Ye:** Data curation (supporting); Methodology (supporting). **Fayong Liu:** Formal analysis (supporting); Methodology (supporting). **Bo Zhou:** Formal analysis (supporting); Methodology (supporting). **Hui Fang:** Conceptualization (supporting); Data curation (supporting); Methodology (supporting). **Runsheng Wang:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Methodology (equal). **Zhigang Ji:** Conceptualization (equal); Data curation (equal);

Formal analysis (equal); Funding acquisition (equal); Supervision (lead); Writing – review & editing (equal). **Ru Huang:** Supervision (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

REFERENCES

[1]A. I. Kingon, J.-P. Maria, and S. K. Streiffer, Nature **406**(6799), 1032 (2000).
[2]C.-H. Lee, S.-H. Hur, Y.-C. Shin, J.-H. Choi, D.-G. Park, and K. Kim, Appl. Phys. Lett. **86**(15), 152908 (2005).
[3]T. Tan, Z. Liu, H. Lu, W. Liu, F. Yan, and W. Zhang, Appl. Phys. A **97**, 475 (2009).
[4]K. Yim, Y. Yong, J. Lee, K. Lee, H.-H. Nahm, J. Yoo, C. Lee, C. Seong Hwang, and S. Han, NPG Asia Mater. **7**(6), e190 (2015).
[5]M. Lee, Y. Youn, K. Yim, and S. Han, Sci. Rep. **8**(1), 14794 (2018).
[6]Y. Umeda, H. Hayashi, H. Moriwake, and I. Tanaka, Jpn. J. Appl. Phys., Part 1 **57**(11S), 11UB01 (2018).
[7]K. Morita, D. W. Davies, K. T. Butler, and A. Walsh, J. Chem. Phys. **153**(2), 024503 (2020).
[8]A. Takahashi, Y. Kumagai, J. Miyamoto, Y. Mochizuki, and F. Oba, Phys. Rev. Mater. **4**(10), 103801 (2020).
[9]E. Kim, J. Kim, and K. Min, Phys. Chem. Chem. Phys. **24**(11), 7050 (2022).
[10]Y. Liu, Z. Yang, X. Zou, S. Ma, D. Liu, M. Avdeev, and S. Shi, Natl. Sci. Rev. **10**(7), nwad125 (2023).
[11]P. Xu, X. Ji, M. Li, and W. Lu, npj Comput. Mater. **9**(1), 42 (2023).
[12]X. Cui, R. Wei, L. Gong, R. Qi, Z. Zhao, H. Chen, K. Song, A. A. Abdulrahman, Y. Wang, J. Z. S. Chen *et al.*, J. Am. Acad. Dermatol. **81**(5), 1176 (2019).
[13]I. O. Alade, M. O. Oyedeji, M. A. A. Rahman, and T. A. Saleh, Soft Comput. **26**(17), 8307 (2022).
[14]M. S. Ahmad, S. M. Adnan, S. Zaidi, and P. Bhargava, Constr. Build. Mater. **248**, 118475 (2020).
[15]S. Wright, J. Agric. Res. **20**(7), 557 (1921).
[16]F. Rustam, A. A. Reshi, A. Mehmood, S. Ullah, B.-W. On, W. Aslam, and G. S. Choi, IEEE Access **8**, 101489 (2020).
[17]J. Nevitt and G. R. Hancock, J. Exp. Educ. **68**(3), 251 (2000).
[18]G. R. Hancock and M. J. Freeman, Educ. Psychol. Meas. **61**(5), 741 (2001).
[19]K. Kelley and K. Lai, Multivar. Behav. Res. **46**(1), 1–32 (2011).
[20]L. Chen, C. Kim, R. Batra, J. P. Lightstone, C. Wu, Z. Li, A. A. Deshmukh, Y. Wang, H. D. Tran, and P. Vashishta, npj Comput. Mater. **6**(1), 61 (2020).
[21]B. Yang, C. R. Zhang, Y. Wang, M. Zhao, H. Y. Yu, Z. J. Liu, X. M. Liu, Y. H. Chen, Y. Z. Wu, and H. S. Chen, Int. J. Quantum Chem. **123**(5), e27039 (2023).
[22]J. Deng and G. Jia, Fluid Phase Equilib. **561**, 113545 (2022).
[23]A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder *et al.*, APL Mater. **1**(1), 011002 (2013).
[24]M. Awad, R. Khanna, M. Awad, and R. Khanna, *Efficient Learning Machines: Theories, Concepts, Applications Engineers System Designers* (Apress, Berkeley, CA, 2015), p. 67.
[25]Y.-J. Lee and S.-Y. Huang, IEEE Trans. Neural Networks **18**(1), 1–13 (2007).
[26]Z. Li, Z. Weida, and J. Licheng, paper presented at the 5th International Conference on Signal Processing Proceedings. 16th World Computer Congress (WCC 2000-ICSP 2000), 2000.
[27]H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, *Advances in Neural Information Processing Systems* (MIT Press, 1996), Vol. 9.
[28]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, J. Mach. Learn. Res. **12**, 2825–2830 (2011).
[29]L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla *et al.*, Comput. Mater. Sci. **152**, 60 (2018).
[30]S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Comput. Mater. Sci. **68**, 314 (2013).
[31]E. Cockayne and B. P. Burton, Phys. Rev. B **62**(6), 3735 (2000).
[32]R. P. Lowndes and D. H. Martin, Proc. R. Soc. A **316**, 351 (1970).
[33]X. Zhao and D. Vanderbilt, Phys. Rev. B **65**(7), 075105 (2002).
[34]M. B. Bashir, M. S. B. Abd Latiff, Y. Coulibaly, and A. Yousif, J. Network Comput. Appl. **60**, 170 (2016).
[35]F. A. Faber, A. Lindmaa, O. A. Von Lilienfeld, and R. Armiento, Phys. Rev. Lett. **117**(13), 135502 (2016).
[36]N. M. Nasrabadi, J. Electron. Imaging **16**(4), 049901 (2007).
[37]L. Breiman, Mach. Learn. **45**, 5 (2001).
[38]M. Nakata and T. Shimazaki, J. Chem. Inf. Model. **57**(6), 1300 (2017).
[39]I. Petousis, D. Mrdjenovich, E. Ballouz, M. Liu, D. Winston, W. Chen, T. Graf, T. D. Schladt, K. A. Persson, and F. B. Prinz, Sci. Data **4**(1), 160134 (2017).
[40]M. Gajdoš, K. Hummer, G. Kresse, J. Furthmüller, and F. Bechstedt, Phys. Rev. B **73**(4), 045112 (2006).
[41]S. Baroni and R. Resta, Phys. Rev. B **33**(10), 7017 (1986).
[42]M. Gao, B. Cai, G. Liu, L. Xu, S. Zhang, and H. Zeng, Phys. Chem. Chem. Phys. **25**(13), 9123 (2023).
[43]S. Sharma, Z. D. Ward, K. Bhimani, M. Sharma, J. Quinton, T. D. Rhone, S.-F. Shi, H. Terrones, and N. Koratkar, ACS Appl. Mater. Interfaces **15**(15), 18962 (2023).
[44]X. YongLin, W. XiangMeng, L. Xin, X. LiLi, N. JianYue, Z. WenHao, W. Zhang, and Y. Jiong, Sci. Sin. Technol. **49**(1), 44 (2018).
[45]W. Jia, Z. Cao, L. Wang, J. Fu, X. Chi, W. Gao, and L.-W. Wang, Comput. Phys. Commun. **184**(1), 9 (2013).
[46]C.-K. Lee, E. Cho, H.-S. Lee, K. S. Seol, and S. Han, Phys. Rev. B **76**(24), 245110 (2007).
[47]W. Hu, L. Zhang, and Z. Pan, ACS Appl. Mater. Interfaces **14**(18), 21596 (2022).
[48]Y. Zhao, J. Zhang, Z. Xu, S. Sun, S. Langner, N. T. P. Hartono, T. Heumueller, Y. Hou, J. Elia, N. Li *et al.*, Nat. Commun. **12**(1), 2191 (2021).
[49]T. Chen, Z. Pang, S. He, Y. Li, S. Shrestha, J. M. Little, H. Yang, T.-C. Chung, J. Sun, H. C. Whitley *et al.*, Nat. Nanotechnol. **19**, 782–791 (2024).
[50]J. Tie and W. Wu, Int. J. Rock Mech. Min. Sci. **179**, 105784 (2024).
[51]Y. Liu, D. Liu, X. Ge, Z. Yang, S. Ma, Z. Zou, and S. Shi, Acta Phys. Sin. **72**(7), 070701 (2023).

Supplementary Materials

Accurate Prediction of Dielectric Properties and Band Gaps in Materials with Machine Learning Approach

Yilin Hu,[1] Maokun Wu,[1,a)] Miaojia Yuan,[1] Yichen Wen,[1] Pengpeng Ren,[1] Sheng Ye,[1] Fayong Liu,[2] Bo Zhou,[3] Hui Fang,[4] Runsheng Wang,[5] Zhigang Ji,[1,a)] and Ru Huang [5]

[1]*National Key Laboratory of Science and Technology on Micro/Nano Fabrication, Shanghai Jiao Tong University, Shanghai 200240, China*

[2]*College of Electronic Engineering, Ocean University of China, Qingdao 266404, China*

[3]*School of Computer Science and Mathematics, Liverpool John Moores University, Byrom Street Liverpool L3 3AF*

[4]*Department of Computer Science, Loughborough University, Loughborough LE11 3TU, U.K*

[5]*School of Integrated Circuits, Peking University, Beijing 100871, China*

Authors to whom correspondence should be addressed:

[a)]E-mail address: maokunwu@sjtu.edu.cn; zhigangji@sjtu.edu.cn

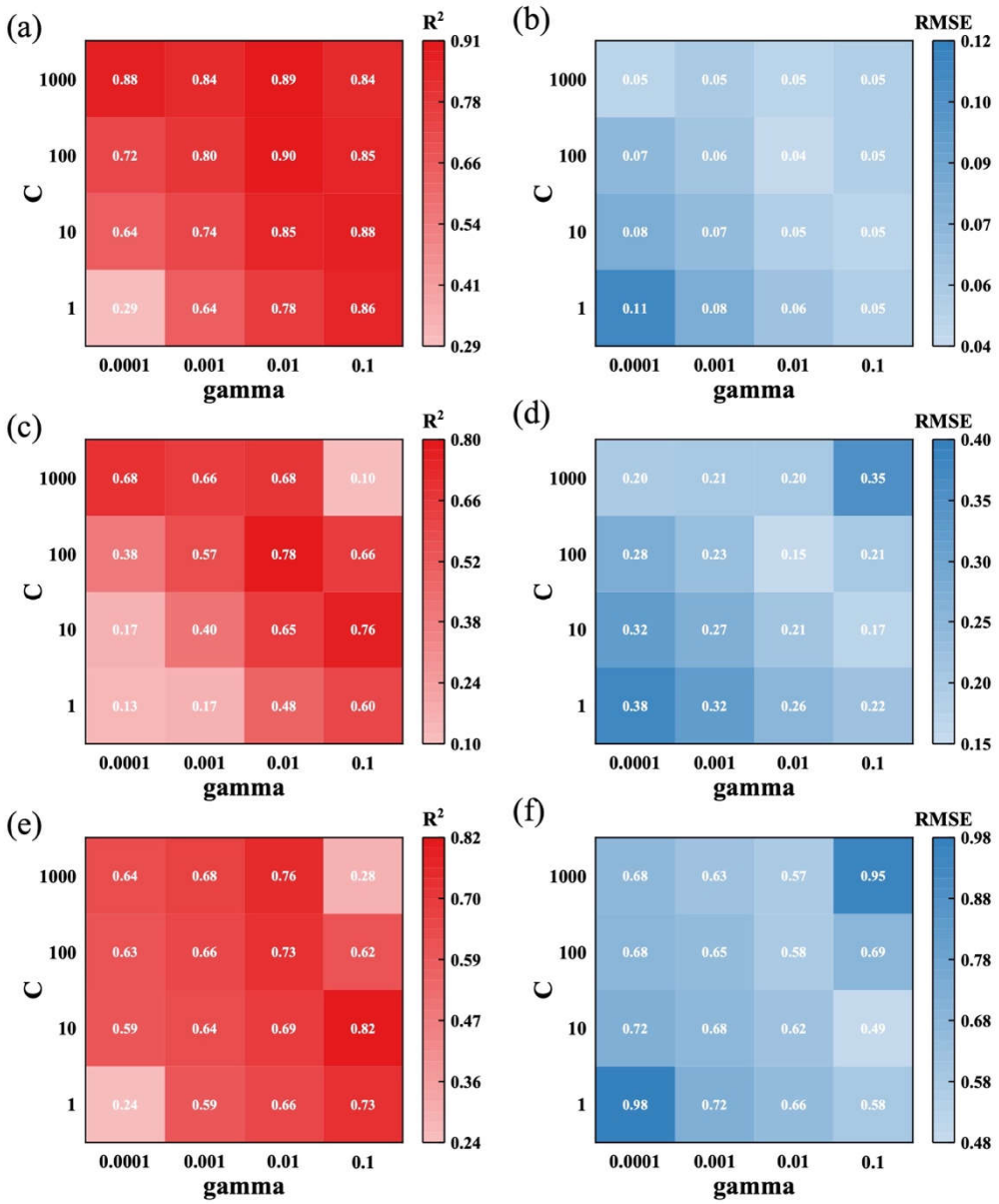

Figure S1. Heatmaps of different hyperparameters C and gamma in the grid search. Red represents the R^2 scores, and blue represents the RMSE. (a) and (b) correspond to the electronic contribution, determined C=100 and gamma=0.01. (c) and (d) show the ionic contribution, determined C=100 and gamma=0.01.(e) and (f) represent the bandgap, determined C=10 and gamma=0.1.

We utilized grid search to explore the hyperparameters C (regularization parameter) and Gamma (kernel coefficient for RBF). For hyperparameter C, we explored values in [1, 10, 100, 1000], and for hyperparameter gamma, we explored values in [0.1, 0.01, 0.001, 0.0001]. By evaluating the model's coefficients of determination (R^2) and Root Mean Square Error

(RMSE), we ultimately determined the optimal hyperparameters as C = (100, 100, 10) and gamma = (0.01, 0.01, 0.1) for the electronic contribution, ionic contribution, and bandgap, respectively. **Figure S1** below shows the heatmap results from the grid search: the red heatmap represents the R^2 results, where deeper red indicates higher fitting coefficients, and the blue heatmap represents the RMSE, where lighter blue indicates smaller errors.