Predicting band gaps and band-edge positions of oxide perovskites using density functional theory and machine learning

Wei Li,[1,2,*] Zigeng Wang,[3,*] Xia Xiao,[3] Zhiqiang Zhang,[4] Anderson Janotti,[1,†]
Sanguthevar Rajasekaran,[3] and Bharat Medasani[5,6,‡]
[1]*Department of Materials Science and Engineering, University of Delaware, Newark, Delaware 19716, USA*
[2]*Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[3]*Computer Science and Engineering Department, University of Connecticut, Storrs, Connecticut 06269, USA*
[4]*Department of Physics, University of Delaware, Newark, Delaware 19716, USA*
[5]*Delaware Energy Institute, University of Delaware, Newark, Delaware 19702, USA*
[6]*Princeton Plasma Physics Laboratory, Princeton, New Jersey 08540, USA*

(Received 9 February 2022; revised 20 September 2022; accepted 4 October 2022; published 28 October 2022)

Density functional theory (DFT) within the local or semilocal density approximations, i.e., the local density approximation (LDA) or generalized gradient approximation (GGA), has become a workhorse in the electronic structure theory of solids, being extremely fast and reliable for energetics and structural properties, yet remaining highly inaccurate for predicting band gaps of semiconductors and insulators. The accurate prediction of band gaps using first-principles methods is time consuming, requiring hybrid functionals, quasiparticle GW, or quantum Monte Carlo methods. Efficiently correcting DFT-LDA/GGA band gaps and unveiling the main chemical and structural factors involved in this correction is desirable for discovering novel materials in high-throughput calculations. In this direction, we use DFT and machine learning techniques to correct band gaps and band-edge positions of a representative subset of ABO_3 perovskite oxides. Relying on the results of HSE06 hybrid functional calculations as target values of band gaps, we find a systematic band-gap correction of ~1.5 eV for this class of materials, where ~1 eV comes from downward shifting the valence band and ~0.5 eV from uplifting the conduction band. The main chemical and structural factors determining the band-gap correction are determined through a feature selection procedure.

DOI: 10.1103/PhysRevB.106.155156

I. INTRODUCTION

The band-gap and band-edge positions (i.e., ionization energy and electron affinity) are basic properties of semiconductors and insulators, and often dictate the suitability of materials for device applications. Their prediction, based on first-principles methods, is key to novel materials discovery. Density functional theory (DFT) calculations [1,2] based on the local density approximation (LDA) [3] or generalized gradient approximation (GGA) [4,5] are often used to predict stable crystal structures, with lattice parameters within 1–2% of the experimental values [6,7]. These calculations are extremely fast and scalable, permitting the study of the energetic and structural properties of thousands of materials with relatively modest computing resources and in relatively short times, playing a central role in current materials discovery research efforts based on high-throughput computation. However, when standard LDA or GGA functionals are employed, band gaps (E_g) predicted by DFT are severely underestimated in comparison to experimental values [8–11]. Predicting E_g of semiconductors and insulators requires going beyond LDA or GGA approximations in DFT, making the calculations much more involved and computationally expensive.

Methods that accurately predict band gaps are very expensive with respect to both computational resources and wall time. The simplest approach is to mix Fock exchange with GGA exchange in a hybrid functional [12–15], partially correcting the self-interaction error in DFT-LDA/GGA, giving band gaps very close to the experimental values for many materials [16–19]. This increases the computation time 10-fold compared to DFT-LDA/GGA calculations. More formally rigorous approaches would be to use the Green's function quasiparticle GW [20–22] or the wave-function-based quantum Monte Carlo [23–25] method, yet at the expense of at least an extra order of magnitude in computational time. As a result, these are not generally amenable to high-throughput computational approaches, posing a stringent obstacle to novel materials discovery.

Machine learning (ML) techniques have emerged as powerful tools in materials science research, with applications in a variety of directions, such as prediction and classification of crystal structures [26–31] and building predictive models of various materials properties [32–35]. Recent efforts also include predicting band gaps, however with limited accuracy [36–40]. A straightforward direction would be to predict band gaps using the DFT-GGA band structures available in the AFLOW database [41] as a training set for machine learning approaches. However, this would have limited use considering that the predicted band gaps would still be severely underestimated. Or one could use DFT+U [42] for band gaps, with

[*] These authors contributed equally to this work.
[†] janotti@udel.edu
[‡] bmedasan@pppl.gov

computational costs similar to those of DFT-LDA/GGA; the problem is what value of U to choose and the justification of applying U to dispersive valence and conduction bands. An interesting approach involves crystal graph convolutional neural networks (CGCNNs) based on atomic connections in the crystal structure after being trained using DFT band gaps [38]. However, this method was also trained and aimed at DFT-GGA band gaps. Recently, reports on automated, high-throughput calculations of band gaps based on a hybrid functional have appeared in the literature [43–46], pointing toward more reliable predictions of band gaps, yet the nature and size of the band-gap corrections from the DFT-GGA values have not been discussed or analyzed.

In this work, we developed machine learning models for mapping band gaps computed with DFT-GGA into band gaps with a higher accuracy Heyd–Scuseria–Ernzerhof (HSE) functional HSE06 hybrid functional. We chose perovskite oxides as an example to demonstrate the applicability of our approach. Oxide perovskites are a class of compounds that are of great importance in technology and basic sciences [47], comprising semiconductors, insulators, ferromagnetic and antiferromagnetic, ferroelectric, multiferroic, piezoelectric, and high-T_c superconductor materials [48]. The wide range of properties is often associated with the orbital character of the bands near the Fermi level and is strongly affected by variations in the crystal structure, such as octahedral rotations and distortions that are associated with deviations from the perfect cubic crystal structure [49]. The accurate prediction of their electronic structure, band gaps, and position of the valence and conduction bands with respect to the vacuum level is crucial for designing novel devices. An interesting feature of ABO_3 perovskite semiconductors and insulators is the dependence of their band gaps on the metal elements A and B, as well as on rotations and tilting of the BO_6 octahedra. Here we restricted the scope of the perovskite materials to those for which the valence band is derived from oxygen $2p$ orbitals and the conduction band is derived from A or B valence orbitals, as indicated in Fig. 1. We did not consider perovskites where the valence and conduction bands are determined by transition-metal d orbitals and the gap associated with spin splitting of d bands or d-d transitions. We explicitly included octahedral tilting and rotations leading to tetragonal, orthorhombic, and rhombohedral crystal structures, as shown in Fig. 1. Using a high-throughput approach [50], we calculated the band structures of the perovskites with Perdew, Burke, and Ernzerhof revised for solids (PBEsol) and HSE06 functionals. We analyzed the mapping of the valence-band maximum (VBM) and conduction-band minimum (CBM) between the PBEsol and HSE06 functionals by employing different machine learning models. Our combined DFT-ML model predicts E_g within an error of 0.16 eV to that of HSE-computed E_g, and reveals the main atomic and structural factors that determine the correction to the VBM, CBM, and, consequently, E_g predicted at the GGA level.

II. METHODS

The first-principles calculations are based on DFT within the generalized gradient approximation of PBEsol [51] and the projector augmented wave method [52,53] as implemented in the Vienna Ab initio Simulation Package (VASP) [54,55]. The wave functions are expanded in plane waves with cutoff energy of 650 eV. Structure optimizations are performed using a 7 × 7 × 7, 7 × 5 × 7, 7 × 5 × 5, and 7 × 7 × 7 Γ-centered k-point grid for the integrations over the Brillouin zones of the cubic, tetragonal, orthorhombic, and rhombohedral primitive cells, respectively. The screened hybrid functional HSE06 [14,15] is employed to compute target band gaps, using the structural parameters found using the PBEsol functional. In tests, we found that PBEsol and HSE06 give lattice parameters that differ by less than 1%, and in good agreement with experimental values. So we neglected the differences in the band gap calculated using the PBEsol-optimized lattice parameters and those calculated using the HSE06-optimized lattice parameters. Test calculations indicate that these differences are less than 0.1 eV.

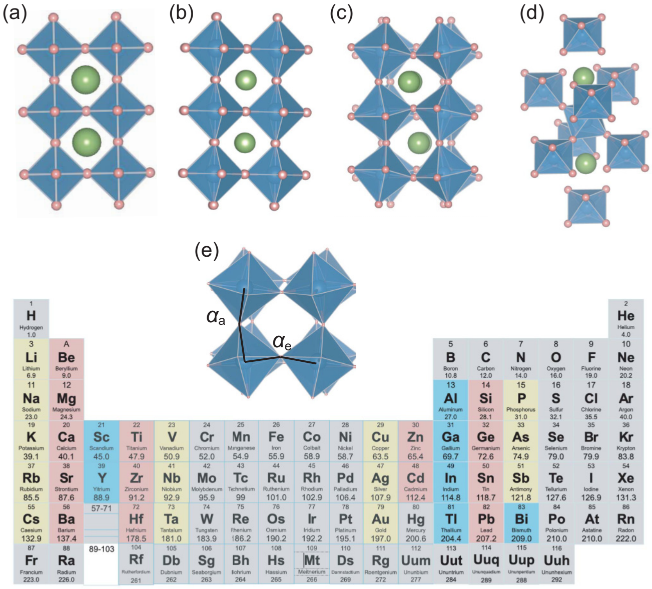
(a)
(b)
(c)
(d)
(e)
α_a
α_e
H He
Li Be B C N O F Ne
Na Mg Al Si P S Cl Ar
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te I Xe
Cs Ba 57-71 Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi Po At Rn
Fr Ra 89-103 Rf Db Sg Bh Hs Mt Ds Rg Uum Uut Uuq Uup Uuh

FIG. 1. Crystal structures of ABO_3 perovskite prototypes and selected A and B atoms. Crystal structure of (a) $Pm\bar{3}m$ cubic, (b) $I4_mmm$ tetragonal, (c) $Pnma$ orthorhombic, and (d) $R\bar{3}c$ rhombohedral structures of ABO_3 perovskites. Green, blue, and red spheres represent A, B, and O atoms, respectively. (e) The apical and equatorial B-O-B bond angles, α_a and α_e. The A and B atoms selected for this study are indicated in the Periodic Table in the lower panel.

We used different ML algorithms to build our band-gap prediction model, including the linear ridge regressor, kernel ridge regressor, and gradient boosted decision tree from open-source software package SCIKIT-LEARN TOOLBOX [56]. The input to the model is comprised of atomic and structural properties, including the B-O-B apical angle α_a and B-O-B equatorial angle α_e. The regression fit to the input gives the predicted band gaps. The prediction performance of the learning models is evaluated by the mean absolute error. The feature importance of all the descriptors is obtained with the gradient boosted decision tree (GBDT) to interpret the importance of various descriptors in the training model. We conducted a hyperparameter search for GBDT models through grid search. The search parameters include max_tree_depth (1, 2,..., 10), number_of_estimators (50, 100, 150,..., 1000),

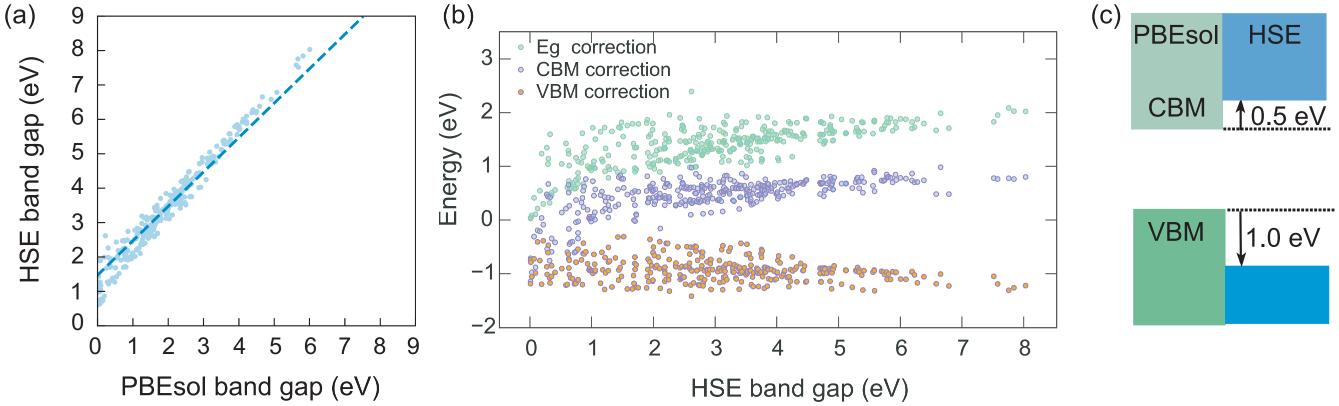

FIG. 2. Correction of the band gap of ABO_3 perovskites based on HSE06 and DFT-GGA PBEsol calculations. (a) HSE06 vs PBEsol band gaps. (b) The band-gap correction (ΔE_g, light green), and correction of the valence-band maximum ΔVBM (light blue) and conduction-band minimum ΔCBM (dark red) vs HSE06 band gap. (c) Schematic of the correction of the band edge of the positions. The dashed line in (a), placed to guide the eye, has slope equal to 1 and crosses the vertical axis at 1.5 eV.

and learning_rate (0.01, 0.02,..., 0.2). We used the default hyperparameter values in the SCIKIT-LEARN package for training linear ridge regressor (LRR) and kernel ridge regressor (KRR). We used MINMAX SCALING to normalize the data for LRR and KRR. We did not normalize the raw features for training GBDT since normalization is not necessary to GBDT due to the tree-based model nature. We partitioned the data such that one-third of the data is reserved for testing. For the remaining two-thirds of the data, threefold cross validation (two-ninths of the total data as the test set and four-tenths of the total data as the training set at any given time) was used for hyperparameter tuning. Our mean absolute error (MAE) results are based on the testing data set.

III. RESULTS AND DISCUSSION

We selected 118 oxide perovskites ABO_3, and for each we considered four crystal structures, with symmetries $Pm\bar{3}m$ (cubic), $I4/mmm$ (tetragonal), $Pnma$ (orthorhombic), and $P63/mmc$ (rhombohedral), as shown in Fig. 1, totaling 472 structures. The selected A and B atoms, also indicated in the Periodic Table in Fig. 1, are A = Li, Na, K, Rb, Cs, Cu, Ag, Au, Be, Mg, Ca, Sr, Ba, Pb, Zn, Cd, Sn, Sc,Y, La, or Bi, and B = P, As, Sb, V, Nb, Ta, Si, Ge, Sn, Ti, Zr, Hf, Al, Ga, In, or Tl, such that the considered compounds satisfy valence(A) + valence(B) = 6. A data set of DFT-GGA band gaps was constructed using this set of materials.

The four crystal structures for all ABO_3 compounds were first optimized with the DFT-GGA PBEsol functional. Then their electronic structures were calculated using PBEsol and HSE06. In this way, since the average electrostatic potential is used as the reference for the Kohn-Sham band energies and does not depend on exchange and correlation, we can directly compare the PBEsol and HSE06 band structures, extracting the corrections for VBM, CBM, and the band gap (i.e., ΔVBM, ΔCBM, and ΔE_g). We note that for all compounds studied here, the VBM for the cubic structure occurs at the R point (0.5, 0.5, 0.5) and the CBM occurs at the Γ point in the cubic Brillouin zone, characterizing an indirect R-Γ fundamental band gap. For the tetragonal, orthorhombic, and rhombohedral structures, both VBM and CBM occur at Γ, characterizing a direct Γ-Γ fundamental band gap.

The calculated HSE06 band gaps vs PBEsol band gaps are shown in Fig. 2(a). There are 383 data points selected in the 472 materials since others are not stable according to the to DFT calculation. First, we note that the HSE06 predicted band gaps have a nearly linear relationship with the DFT-GGA predicted band gaps. We applied a simple linear regression fit using $y = ax + b$ between the two sets of band gaps and obtained $a = 1.12$ and $b = 1.15$. The resulting mean absolute error (MAE) is 0.21 eV, which is comparable to the MAEs obtained with the more complicated models presented in the study. Since the value of a is close to 1, the data had been fit to an even simpler model of fixed correction, $y = x + b'$. Fixed correction is very appealing due to its simplicity and provides an intuitive physical insight into the nature of the correction. The optimal b' was found to be 1.5 eV with an MAE of 0.32 eV. The MAE of the fixed correction model compares well with the typical error in the DFT predicted band gaps, even when hybrid functionals are used. The fixed correction model implies that DFT-GGA underestimates the band gap with respect to HSE06 by ~1.5 eV. This is quite surprising given that in general, DFT-LDA/GGA does not underestimate the band gap of semiconductors and insulators by a fixed amount [57]. The largest deviation from this trend is observed for compounds containing Cu, Pb, and Sn occupying the A site. In the case of Cu-B-O_3 compounds, the Cu d orbitals mix with the O $2p$ orbitals, pushing the VBM to higher energies. In the case of Sn-B-O_3 and Pb-B-O_3, the VBM has large contributions from Sn and Pb s valence orbitals, which also pushes the VBM to higher energies. In all the cases where the valence band is mostly derived from O $2p$ orbitals, the approximate 1.5 eV band-gap correction fits the data quite well.

The separated corrections ΔVBM and ΔCBM, i.e., the amount the VBM and CBM in HSE06 differ from the VBM and CBM in DFT-GGA, are shown in Fig. 2(b). Contrary to

TABLE I. Mean absolute error (MAE) used to evaluate the performance of fixed correction (FC), linear regression (LR), linear ridge regressor (LRR), kernel ridge regressor (KRR), and the gradient boosted decision tree (GBDT) models in predicting the corrections of the valence-band maximum (ΔVBM), conduction-band minimum (ΔCBM), and band gap (ΔE_g) of oxide perovskites in DFT-GGA PBEsol compared to the HSE06 values.

	FC	LR	LRR	KRR	GBDT
ΔVBM	0.17	0.09	0.10 ± 0.01	0.09 ± 0.01	0.09 ± 0.01
ΔCBM	0.24	0.17	0.19 ± 0.01	0.15 ± 0.01	0.10 ± 0.003
ΔE_g	0.32	0.21	0.23 ± 0.02	0.20 ± 0.01	0.16 ± 0.01

common wisdom, where it is often assumed that to correct the DFT-GGA band gap only, an upward shift of the CBM is necessary, we find that about 2/3 of the gap correction comes from shifting down the VBM and only about 1/3 of the correction comes from shifting the conduction band upward. This is attributed to large self-interaction correction of the O 2*p*-derived valence bands in these materials. Again, the outliers, where the VBM is corrected by a lesser amount, correspond to compounds containing Cu, Sn, or Pb in the *A* site. It is also interesting to note that the correction in the VBM derived from O 2*p* is larger than the correction of CBM derived from *d* orbitals, such as in $SrTiO_3$ and similar compounds, despite the rather flat nature of their conduction bands that are derived from the quite localized transition-metal *d* orbitals. Finally, we also note that the band-gap correction ΔE_g is slightly larger than 1.5 eV for compounds with larger band gaps, approaching 2 eV, and this is traced back to the correction of the CBM which approaches 1 eV for compounds with $E_g \gtrsim 4$ eV.

Having established the band-gap correction for these oxide perovskites, we now turn to machine learning techniques to develop a model that correlates the ΔVBM, ΔCBM, and ΔE_g corrections to atomic and structural properties of the compounds. The atomic properties as input to the machine learning models include electronegativity, ionization energy, valence-orbital energies, and atomic radius of both *A* and *B* atoms. Structural properties include octahedral tilting and rotations that are characterized by the apical α_a and equatorial α_e angles corresponding to *B*-O-*B* angles parallel and perpendicular to the *c* axis. We employed three machine learning models, which are the linear ridge regressor (LRR), kernel ridge regressor (KRR), and the gradient boosted decision tree (GBDT) regressor, as implemented in the SCIKIT-LEARN TOOLBOX [56]. We used a regularization strength of 0.01 to both LRR and KRR models. For the KRR method, we used a polynomial kernel with a maximum order of 3. For the GBDT model, we set the maximum tree depth to 5 with 500 base estimators.

The prediction performance of the LRR, KRR, and GBDT models can be seen in Table I. In these models, we use two-thirds of the data as the training set. We also use the mean absolute error (MAE) to measure the performance in predicting ΔVBM, ΔCBM, and ΔE_g. Among the three models, GBDT gives the highest prediction accuracy with low variance; the KRR model performs better than LRR. Note that we obtain lower MAE than previous models [31,36,44,58,59], likely due to the better quality or more uniformity of our training dataset. The results indicate that there exists a nonlinear relation between the input properties and the target results, explaining why the pure linear model LRR performs poorly. Note that all three ML models predict ΔVBM with similar performance, indicating that the VBM correction has a more linear relationship with the input properties than the CBM and E_g corrections.

What are the main atomic and structural properties that determine the band-gap and band-edge corrections? The answer is shown in Fig. 3, where the input properties are ranked according to their contributions to the prediction accuracy based on the mean decrease in the impurity of the GBDT model [60]. We find that the electronegativity, the energy of the *p* valence orbital of atom *A*, and the equatorial angle of the octahedral rotation are the main properties that determine ΔVBM. For ΔCBM, the main properties are the electronegativity, ionization energy of atom *B*, and the equatorial angle of the octahedral rotation. For more advanced feature importance evaluation methods with higher local and global consistency and interpretability, we refer to the literature [61,62]. We have also applied LRR, KRR, and GBDT models to the data by excluding the discovered less-important features for each label, and no obvious accuracy improvement is identified.

For both ΔVBM and ΔCBM, the equatorial angle determines the overlap between the orbitals of *B* and O in the directions parallel to the *a*-*b* plane, which, in turn, affect both the VBM and CBM positions. Note that the dependence on the apical angle α_a is less than that on the equatorial angle α_e since the former affects the *B*-O orbital overlap only along the *c* direction. Finally, we also note that the relative importance of the electronegativity, ionization energies, and rotation angles is higher for atom *A* than for atom *B* in determining the band gap. This is attributed to the larger contribution of the VBM correction than the CBM correction to ΔE_g.

IV. SUMMARY

Using high-throughput DFT-GGA PBEsol and HSE06 calculations, we determined the band-gap correction of a representative set of oxide perovskites, finding that the HSE06-based correction pushes down the valence band by ~1 eV and pushes up the conduction band by ~0.5 eV. These results are then used in machine learning models that include atomic and structural properties as input to determine the corrections to the valence band, conduction band, and band gap. The properties used as fitting parameters are ranked according to their relative importance to the corrections. We find that the electronegativity of the *A* and *B* atoms together with the equatorial angle of rotation of the BO_6 octahedra are the main factors involved in the corrections. These results serve as a starting point and guide to developing machine-learning-based approaches applicable to the discovery of novel electronic materials.

The datasets generated and/or analyzed during the current study are available in the GitHub repository [63].

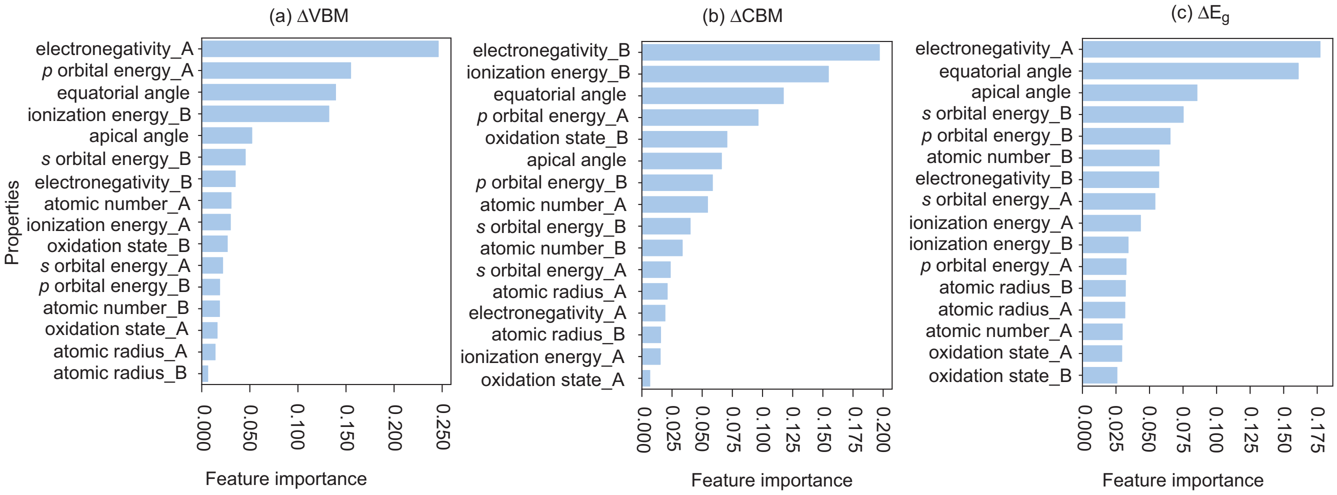

FIG. 3. Feature importance in the gradient boosted decision tree (GBDT) model for determining the band-gap (ΔE_g) and band-edge (ΔVBM, ΔCBM) corrections of ABO_3 perovskites.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation Grant No. EAGER-1843025. This research was also supported by the Xtreme Science and Engineering Discovery Environment (XSEDE) facility, National Science Foundation Grant No. ACI-1053575, and the Information Technologies (IT) resources at the University of Delaware, specifically the high-performance computing resources. This research was also supported in part by National Science Foundation Grants No. 1447711, No. 1743418, and No. 1843025. A.J. acknowledges support from NSF Faculty Early Career Development Program Grant No. DMR-1652994. W.L. was partly supported by the Laboratory Directed Research and Development Program of Los Alamos National Laboratory (LANL) under Project No. 20210087DR. LANL is operated by Triad National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy (Contract No. 89233218CNA000001).

[1] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[2] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[3] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
[4] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[5] J. P. Perdew, W. Yang, K. Burke, Z. Yang, E. K. U. Gross, M. Scheffler, G. E. Scuseria, T. M. Henderson, I. Y. Zhang, A. Ruzsinszky, H. Peng, J. Sun, E. Trushin, and A. Görling, Proc. Natl. Acad. Sci. **114**, 2801 (2017).
[6] F. Tran, J. Stelzl, and P. Blaha, J. Chem. Phys. **144**, 204120 (2016).
[7] L. He, F. Liu, G. Hautier, M. J. T. Oliveira, M. A. L. Marques, F. D. Vila, J. J. Rehr, G.-M. Rignanese, and A. Zhou, Phys. Rev. B **89**, 064305 (2014).
[8] J. P. Perdew, Intl. J. Quantum Chem. **28**, 497 (1985).
[9] J. P. Perdew and M. Levy, Phys. Rev. Lett. **51**, 1884 (1983).
[10] L. J. Sham and M. Schlüter, Phys. Rev. Lett. **51**, 1888 (1983).
[11] P. Mori-Sánchez, A. J. Cohen, and W. Yang, Phys. Rev. Lett. **100**, 146401 (2008).
[12] A. D. Becke, J. Chem. Phys. **98**, 1372 (1993).
[13] J. P. Perdew, M. Ernzerhof, and K. Burke, J. Chem. Phys. **105**, 9982 (1996).
[14] J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. **118**, 8207 (2003).
[15] J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. **124**, 219906 (2006).
[16] E. N. Brothers, A. F. Izmaylov, J. O. Normand, V. Barone, and G. E. Scuseria, J. Chem. Phys. **129**, 011102 (2008).
[17] H. Xiao, J. Tahir-Kheli, and W. A. Goddard, J. Phys. Chem. Lett. **2**, 212 (2011).
[18] Y.-S. Kim, K. Hummer, and G. Kresse, Phys. Rev. B **80**, 035203 (2009).
[19] T. M. Henderson, J. Paier, and G. E. Scuseria, Phys. Stat. Sol. (b) **248**, 767 (2011).
[20] M. S. Hybertsen and S. G. Louie, Phys. Rev. B **34**, 5390 (1986).
[21] M. Shishkin and G. Kresse, Phys. Rev. B **74**, 035101 (2006).
[22] W. Chen and A. Pasquarello, Phys. Rev. B **92**, 041115(R) (2015).
[23] R. J. Hunt, M. Szyniszewski, G. I. Prayogo, R. Maezono, and N. D. Drummond, Phys. Rev. B **98**, 075122 (2018).
[24] Y. Yang, V. Gorelov, C. Pierleoni, D. M. Ceperley, and M. Holzmann, Phys. Rev. B **101**, 085115 (2020).
[25] R. J. Hunt, B. Monserrat, V. Zólyomi, and N. D. Drummond, Phys. Rev. B **101**, 205115 (2020).
[26] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, Nat. Mater. **5**, 641 (2006).
[27] D. A. Carr, M. Lach-hab, S. Yang, I. I. Vaisman, and E. Blaisten-Barojas, Micropor. Mesopor. Mater. **117**, 339 (2009).
[28] G. Pilania, J. E. Gubernatis, and T. Lookman, Phys. Rev. B **91**, 214302 (2015).
[29] T. Yamashita, N. Sato, H. Kino, T. Miyake, K. Tsuda, and T. Oguchi, Phys. Rev. Mater. **2**, 013803 (2018).

[30] W. Ye, C. Chen, Z. Wang, I.-H. Chu, and S. P. Ong, Nat. Commun. **9**, 3800 (2018).

[31] A. C. Rajan, A. Mishra, S. Satsangi, R. Vaish, H. Mizuseki, K.-R. Lee, and A. K. Singh, Chem. Mater. **30**, 4031 (2018).

[32] B. Medasani, A. Gamst, H. Ding, W. Chen, K. A. Persson, M. Asta, A. Canning, and M. Haranczyk, Npj Comput. Mater. **2**, 1 (2016).

[33] J. Lee, A. Seko, K. Shitara, K. Nakayama, and I. Tanaka, Phys. Rev. B **93**, 115104 (2016).

[34] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, and T. Lookman, Sci. Rep. **6**, 19375 (2016).

[35] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, Sci. Adv. **3**, e1701816 (2017).

[36] Y. Zhuo, A. M. Tehrani, and J. Brgoch, J. Phys. Chem. Lett. **9**, 1668 (2018).

[37] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, and J. Wang, Nat. Commun. **9**, 3405 (2018).

[38] T. Xie and J. C. Grossman, Phys. Rev. Lett. **120**, 145301 (2018).

[39] O. Allam, C. Holmes, Z. Greenberg, K. C. Kim, and S. S. Jang, Chem. Phys. Chem. **19**, 2559 (2018).

[40] B. Olsthoorn, R. M. Geilhufe, S. S. Borysov, and A. V. Balatsky, Adv. Quantum Technol. **2**, 1900023 (2019).

[41] S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, and D. Morgan, Comput. Mater. Sci. **58**, 218 (2012).

[42] V. I. Anisimov, F. Aryasetiawan, and A. I. Lichtenstein, J. Phys.: Condens. Matter **9**, 767 (1997).

[43] J. Jie, M. Weng, S. Li, D. Chen, S. Li, W. Xiao, J. Zheng, F. Pan, and L. Wang, Sci. China Technol. Sci. **62**, 1423 (2019).

[44] G. Pilania, J. E. Gubernatis, and T. Lookman, Comput. Mater. Sci. **129**, 156 (2017).

[45] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, Npj Comput. Mater. **2**, 16028 (2016).

[46] Y. Huang, C. Yu, W. Chen, Y. Liu, C. Li, C. Niu, F. Wang, and Y. Jia, J. Mater. Chem. C **7**, 3238 (2019).

[47] H. Wenk and A. Bulakh, *Minerals: Their Constitution and Origin* (Cambridge University Press, New York, 2004).

[48] P. Szuromi and B. Grocholski, Science **358**, 732 (2017).

[49] M. A. Peña and J. L. G. Fierro, Chem. Rev. **101**, 1981 (2001).

[50] A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, D. Gunter, and K. A. Persson, Concurrency Computat.: Pract. Expt. **27**, 5037 (2015).

[51] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, Phys. Rev. Lett. **100**, 136406 (2008).

[52] P. E. Blöchl, Phys. Rev. B **50**, 17953 (1994).

[53] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).

[54] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).

[55] G. Kresse and J. Hafner, Phys. Rev. B **48**, 13115 (1993).

[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Res **12**, 2825 (2011).

[57] Y. Hinuma, A. Grüneis, G. Kresse, and F. Oba, Phys. Rev. B **90**, 155405 (2014).

[58] V. Gladkikh, D. Y. Kim, A. Hajibabaei, A. Jana, C. W. Myung, and K. S. Kim, J. Phys. Chem. C **124**, 8905 (2020).

[59] A. Mishra, S. Satsangi, A. C. Rajan, H. Mizuseki, K.-R. Lee, and A. K. Singh, J. Phys. Chem. Lett. **10**, 780 (2019).

[60] L. Breiman, Mach. Learn. **45**, 5 (2001).

[61] S. M. Lundberg and S.-I. Lee, *Advances in Neural Information Processing Systems*, Vol. 30 (Curran Associates, Inc., New York, 2017).

[62] S. M. Lundberg, G. G. Erion, and S.-I. Lee, arXiv:1802.03888.

[63] See https://github.com/vera-weili/perovskite_ML.