



CHORUS

This is the accepted manuscript made available via CHORUS. The article has been published as:

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

Wei Li, Zigeng Wang, Xia Xiao, Zhiqiang Zhang, Anderson Janotti, Sanguthevar Rajasekaran, and Bharat Medasani

Phys. Rev. B **106**, 155156 — Published 28 October 2022

DOI: [10.1103/PhysRevB.106.155156](https://doi.org/10.1103/PhysRevB.106.155156)

Predicting bandgaps and band-edge positions of oxide perovskites using DFT and machine learning

Wei Li,^{1,2} Zigeng Wang,³ Xia Xiao,³ Zhiqiang Zhang,⁴ Anderson
Janotti,¹ Rajasekaran Sanguthevar,³ and Bharat Medasani^{5,6}

¹*Department of Materials Science & Engineering,
University of Delaware, Newark, DE 19716, USA*

²*Computer, Computational, and Statistical Sciences Division,
Los Alamos National Laboratory, Los Alamos, NM, 87545, USA*

³*Computer Science and Engineering Department,
University of Connecticut, CT 06269, USA*

⁴*Department of Physics, University of Delaware, Newark, DE 19716, USA*

⁵*Delaware Energy Institute, University of Delaware, Newark, DE 19702, USA*

⁶*Princeton Plasma Physics Laboratory, Princeton, NJ 08540**

Abstract

Density functional theory within the local or semilocal density approximations (DFT-LDA/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 bandgaps of semiconductors and insulators. Accurate prediction of bandgaps using first-principles methods is time-consuming, requiring hybrid functionals, quasi-particle GW, or quantum Monte Carlo methods. Efficiently correcting DFT-LDA/GGA bandgaps 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 bandgaps 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 bandgaps, we find a systematic bandgap 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 bandgap correction are determined through a feature selection procedure.

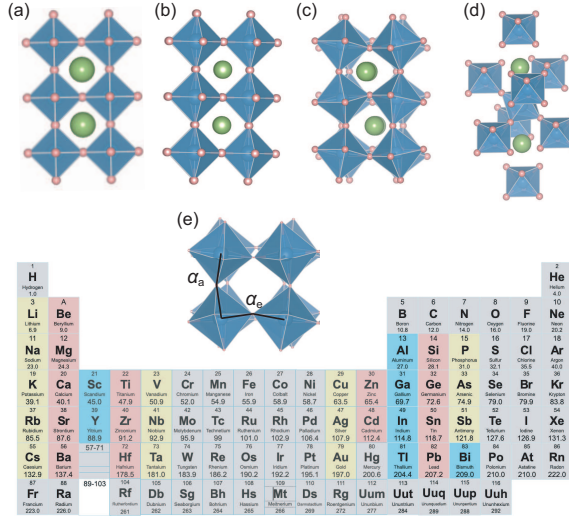


FIG. 1. Crystal structures of ABO_3 perovskite prototypes and selected A and B atoms. Crystal structure of (a) $Pm\bar{3}m$ cubic, (b) $I4mm$ tetragonal, (c) $Pnma$ orthorhombic, and (d) $R\bar{3}c$ rhombohedral structures of ABO_3 perovskites. Green, blue and red spheres represent A, B and O atoms, respectively. The apical and equatorial B-O-B bond angles, α_a , and α_e are indicated in (e). The A and B atoms selected for this study are indicated in the Periodic Table in the lower panel.

INTRODUCTION

The bandgap 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. DFT calculations^{1,2} based on LDA³ or 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, bandgaps (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 bandgaps 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 bandgaps very close to the experimental values for many materials.^{16–19} This increases the computation time tenfold compared to DFT-LDA/GGA calculations. More formally rigorous approaches would be to use the Greens function quasiparticle GW^{20–22} or the wavefunction-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 bandgaps, however with limited accuracy.^{36–40} A straightforward direction would be to predict bandgaps using the DFT-GGA band structures available in AFLOW database⁴¹ as a training set for machine learning approaches. However, this would have limited use considering that the predicted bandgaps would still be severely underestimated. Or one could use DFT+U⁴² for bandgaps, with 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 graph convolutional neural networks (CGCNN) based on atomic connections in the crystal structure after being trained using DFT bandgaps.³⁸ However, this method was also trained and aimed at DFT-GGA bandgaps. Recently, reports on automated, high-throughput calculations of bandgaps based on hybrid functional have appeared in the literature,^{43–46} pointing toward more reliable predictions of bandgaps, 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 bandgaps computed with DFT-GGA into bandgaps with higher accuracy HSE06 hybrid functional. We chose perovskite oxides as 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

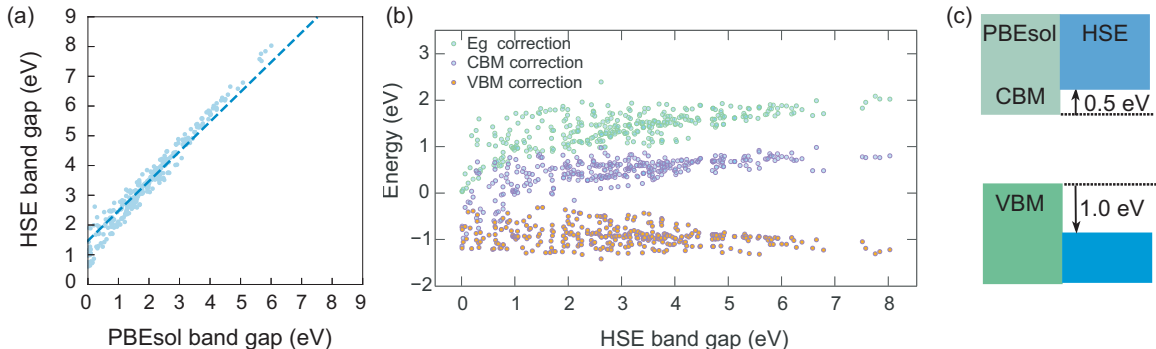


FIG. 2. Correction of the bandgap of ABO_3 perovskites based on HSE06 and DFT-GGA PBEsol calculations. (a) HSE06 vs PBEsol bandgaps, (b) the band-gap correction (ΔE_g , light green), correction of the valence-band maximum ΔVBM (light blue) and conduction-band minimum ΔCBM (dark red) vs HSE06 bandgap. (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 1 and crosses the vertical axis at 1.5 eV.

level and are 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 Accurate prediction of their electronic structure, bandgaps, and position of valence and conduction bands with respect to 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 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 PBEsol and HSE06 functionals. We analyzed the mapping of valence band maximum (VBM) and conduction band minimum (CBM) between 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 GGA level.

I. METHODS

The first-principles calculations are based on DFT within the generalized gradient approximation of Perdew, Burke, and Ernzerhof revised for solids (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 $7\times 7\times 7$, $7\times 5\times 7$, $7\times 5\times 5$, and $7\times 7\times 7$ 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 bandgaps, 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 bandgap 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.

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 bandgaps. The prediction performance of the learning models is evaluated by the mean absolute error. The feature importance of all the descriptors is obtained with GBDT to interpret the importance of various descriptors in the training model. We conducted a hyper-parameter search for GBDT models through grid search. The search parameters include `max_tree_depth` (1, 2, ... 10), `number_of_estimators` (50, 100, 150, ..., 1000), and `learning_rate` (0.01, 0.02, ..., 0.2). We used the default hyper-parameter values in scikit-learn package for training LRR and 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

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 bandgap (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

reserved for testing. For the remaining two-thirds of the data, three-fold cross-validation (two-ninths of total data as the test set and four-tenths of total data as the training set at any given time) was used hyper-parameter tuning. Our MAE results are based on the testing data set.

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 $\text{valence(A)} + \text{valence(B)} = 6$. A data set of DFT-GGA bandgaps 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 and CBM, and the bandgap (i.e., ΔVBM , ΔCBM , and ΔE_g). We note that for all

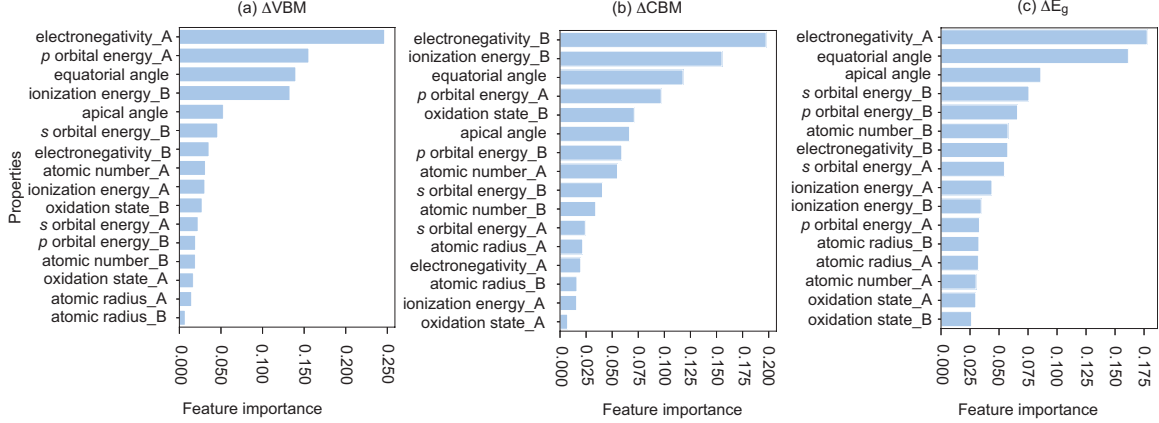


FIG. 3. Feature importance in the gradient boosted decision tree (GBDT) model for determining the band-gap (ΔE_g) and band-edge corrections (ΔVBM , ΔCBM) of ABO_3 perovskites.

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 bandgap. For the tetragonal, orthorhombic, and rhombohedral structures, both VBM and CBM occur at Γ , characterizing a direct Γ - Γ fundamental bandgap.

The calculated HSE06 bandgaps vs PBEsol bandgaps are shown in Fig. 2(a) and Supplementary Information. There are 383 data selected in the 472 materials since others are not stable according to the DFT calculation. First, we note that the HSE06 predicted bandgaps have a nearly linear relationship with the DFT-GGA predicted bandgaps. We applied a simple linear regression fit using $y = ax + b$ between the two sets of bandgaps, and obtained $a = 1.12$ and $b = 1.15$. The resulting mean absolute error (MAE) is 0.21 eV, which is comparable to the MAE's 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 bandgaps even when hybrid functionals are used. The fixed correction model implies that DFT-GGA underestimates the bandgap with respect to HSE06 by ~ 1.5 eV. This is quite surprising given that in general DFT-LDA/GGA does not underestimate bandgap 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-O₃ compounds, the Cu *d* orbitals mix with the O *2p* orbitals, pushing the VBM to higher energies. In the case of Sn-B-O₃ and Pb-B-O₃, 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 common wisdom, where it is often assumed that to correct the DFT-GGA bandgap 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 *2p*-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 the correction in the VBM derived from O *2p* is larger than the correction of CBM derived from *d* orbitals, such as in SrTiO₃ 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 bandgaps, approaching 2 eV, and this is traced back to the correction of the CBM which approaches 1 eV for compounds with E_g is $\gtrsim 4$ 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 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 third of the data as the training set. We also use 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 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 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 atom 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 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 bandgap. This is attributed to the larger contribution of the VBM correction than the CBM correction to ΔE_g .

SUMMARY

Using high-throughput DFT-GGA PBEsol and HSE06 calculations we determined the bandgap 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 bandgap. The properties used as fitting parameters are ranked according to their relative importance to the corrections. We find that 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 starting point and guide to developing machine-learning-based approaches applicable to the discovery of novel electronic materials.

DATA AVAILABILITY

The datasets generated and/or analysed during the current study are available in the GitHub repository https://github.com/vera-weili/perovskite_ML

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation EAGER-1843025. This research was also supported by the Xtreme Science and Engineering Discovery Environment (XSEDE) facility, National Science Foundation grant number 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 the following National Science Foundation grants: 1447711, 1743418, and 1843025. AJ acknowledges support from NSF Faculty Early Career Development Program DMR-1652994. Wei Li was partly supported by the Laboratory Directed Research and Development Program of Los Alamos National Laboratory (LANL) under project number 20210087DR. LANL is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S.

- * bmedasan@pppl.gov
- ¹ P. Hohenberg and W. Kohn, Phys. Rev. B **7**, 1912 (1964).
 - ² W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
 - ³ P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
 - 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. Grling, 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, Int. 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. Status Solidi 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 (2015).

- ²³ R. J. Hunt, M. Szytniszewski, 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, *Nature Materials* **5**, 641 (2006).
- ²⁷ D. A. Carr, M. Lach-hab, S. Yang, I. I. Vaisman, and E. Blaisten-Barojas, *Microporous Mesoporous 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. Hanczyk, *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** (2017).
- ³⁶ Y. Zhuo, M. T. A, 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, *ChemPhysChem* **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. Condens. Matter Phys.* **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, 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, *Concurr. Comput.* **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, *Machine learning* **45**, 5 (2001).
- ⁶¹ S. M. Lundberg and S.-I. Lee, *Advances in neural information processing systems* **30** (2017).
- ⁶² S. M. Lundberg, G. G. Erion, and S.-I. Lee, *arXiv preprint arXiv:1802.03888* (2018).