THE JOURNAL OF PHYSICAL CHEMISTRY Letters

Surfaces, Interfaces, and Catalysis; Physical Properties of Nanomaterials and Materials

Predicting the Band Gaps of Inorganic Solids by Machine Learning

Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch

J. Phys. Chem. Lett., **Just Accepted Manuscript** • DOI: 10.1021/acs.jpclett.8b00124 • Publication Date (Web): 13 Mar 2018

ACS Publications

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Predicting the Band Gaps of Inorganic Solids by Machine Learning

*Ya Zhuo, Aria Mansouri Tehrani, Jakoah Brgoch**
Department of Chemistry, University of Houston, Houston, Texas 77204, United States

ABSTRACT: A machine-learning model is developed that can accurately predict the band gap of inorganic solids based only on composition. This method uses support vector classification to first separate metals from non-metals, followed by quantitatively predicting the band gap of the non-metals using support vector regression. The superb accuracy of the regression model is obtained by using a training set is composed entirely of experimentally measured band gaps and utilizing only compositional descriptors. In fact, because of the unique training set of experimental data, the machine learning predicted band gaps are significantly closer to the experimentally reported values than DFT (PBE-level) calculated band gaps. This resulting tool not only provides the ability to accurately predict the band gap for any composition, but the versatility and speed of the prediction based only on composition will make this a great resource to screen inorganic phase space and direct the development of functional inorganic materials.

Functional inorganic materials that are used in a myriad of applications like LEDs[1,2], transistors[3], photovoltaics[4,5], or scintillators[6] require meticulous knowledge of the band gap (E_g). To accelerate the development of new materials for these applications, density functional theory

(DFT) is often used to predict E_g *a priori* and direct synthetic efforts toward optimal materials.[7] As a result, high-throughput computation has made tremendous strides in determining the electronic structures for thousands of inorganic solids.[8,9] Nevertheless, limitations to calculating the electronic structure of solids using conventional DFT remain.[10,11] Most notably, there is a systematic underestimation of E_g compared to the experimental values when employing standard exchange and correlation functionals like PBE. Beyond conventional DFT, methods such as hybrid functionals[12,13] or GW-type methods significantly improve the calculation of E_g;[14] however, these methods are not currently amenable to high-throughput computation due to their excessive computational cost. Alternatively, implementation of functionals such as Becke-Johnson (mBJ)[15,16] or the generalization of Δ-SCF to solids (Δ-sol)[17] provide an accurate prediction of band gaps without a significant increase of computational cost. The former demonstrates excellent performance for most semiconductors and insulators but suffers in describing ferromagnetic metals[18] whereas the latter depends on the dielectric screening properties of electrons and is not parameter free.[17,19] As a result, these are not generally implemented in the high-through frameworks. DFT is also often limited to ordered crystal structures and is not reliable for highly correlated systems, although DFT+U is able to improve the outcome.[20] Considering ≈16% of Pearson's Crystal Structure Database (PCD) contains atomic disorder such as statistical mixing and ≈23% of PCD contains rare-earth atoms, a different approach is required to predict E_g for entire databases of compounds or general compositions.

Research has recently turned to applying machine learning as a way of circumventing many of these limitations as well as supplement DFT for estimating band gaps. For example, a cross-validated kernel ridge regression (KRR) model can reliably predict the band gap of double

perovskites.[21] A method based on a neural network method was also developed to predict E_g for chalcopyrite as well group III–V and II–VI binary[22] and group I–III–VI2 and II–IV–V2 ternary semiconductors.[23] Artificial neural networks[24] or support vector regression[25] have also been used to predict the band gap of binary and ternary semiconductors, although, these employed a small training set (<30) of compounds. Finally, a multi-fidelity Gaussian process (GP) based co-kriging regression model was utilized to predict E_g for hundreds of elpasolite compounds.[19] Expanding beyond on specific crystal systems, a more general method has been established that used the calculated electronic structure available within the AFLOW project[26] as a training set for machine learning. This approach uses universal fragment descriptors with a gradient boosting decision tree to provide robust estimation not only of the band gap for inorganic solids, but also for other thermochemical properties like heat capacity, Debye temperature, and the elastic moduli.[27] Similarly, another approach that uses crystal graph convolutional neural networks based on atomic connections in the crystal can reach the accuracy of DFT calculations after being trained using DFT band gaps.[28] These methods are all viable for quickly estimating properties like E_g. Nevertheless, these machine-learning models are all trained using E_g values that are (inaccurately) determined by PBE-level DFT or DFT+U, which limits their ability to replicate the experimental band gap values unless additional empirical corrections are included.

Here, we report the development of a supervised machine-learning approach to predict the E_g of inorganic materials without relying on DFT calculations. Our training set employs 3896 experimentally reported band gaps comprised of 2458 unique compositions obtained from diffuse reflectance, resistivity measurements, surface photovoltage, photo conduction, and UV/Vis measurements. The complete training set of experimental data is provided as Supporting Information. The advantage of this approach is that by employing a set of experimentally

measured E_g values available in the literature,[29–32] our machine-learning predictions are not subject to the same systematic error as DFT determined band gaps. Moreover, the resulting machine-learning model is capable of predicting the band gap using a descriptor set based only on the elemental properties of the constituent elements, which are related to the atom's relative position on the periodic table, the electronic structure, and its physical properties, among other descriptors. The full list of variables is provided in Supporting Information Table S1. The descriptor set is limited to composition descriptors in our current machine-learning model because most of the band gaps obtained from the literature are not accompanied by sufficient crystallographic data, only the necessary composition information. Nonetheless, the number of descriptors can be expanded by describing the compositions through several mathematical formulae involving simple combinations of these elemental properties (e.g., sum, difference) or their extremes (e.g., largest or smallest values). To minimize the detriment from excessive matrix dimension, a genetic algorithm-based feature selection using partial least square method was conducted. Interestingly, reducing the number of features didn't improve the outcome, so all features were included in the final model. The advantage of using a training set derived from experimentally measured band gaps and a composition-based descriptor set is that the machine-learning method can rapidly and accurately predict the band gap for any given composition.

The supervised machine-learning method occurs in two stages. First, a classification was conducted to classify a compound as metal ($E_g = 0$ eV) or non-metal ($E_g > 0$ eV). The band gaps of the compounds classified as non-metals were then quantitatively determined using a regression method. Several regressors and classifiers including support vector machine (with linear, polynomial, and radial basis function kernels), K-Nearest Neighbors (KNN), kernel ridge regression (KRR), and logistic regression (LG) were tested. These results are provided in

Supporting Information Table S2. The model using support vector machine with a radial basis function (RBF) kernel resulted in the best performance.

The support vector classification (SVC)[33,34] that was employed analyzed an equal number of metals, selected based on their 0 eV E_g calculated by the Materials Project,[35] and non-metals using the experimentally measured band gaps extracted from the chemical literature. Balancing the dataset is essential to avoid bias in the eventual predictions. The total dataset contained 4916 compositions, which was then divided in to a training set (80%) and a test set (20%).

Using the SVC approach, the probability for each composition in the test set is predicted to be a metal (0-0.5) or non-metal (0.5-1). As shown in Figure 1a, it is apparent that a majority of the compositions are well segregated into each class with a small spread across the probability range. In fact, only 8% of the compounds (79 out of 984) are misclassified for the test set (8.6% for cross-validated training set). These statistics confirm the model is able to reliably segregate metals from non-metals with a few exceptions, most of which have unusual compositions or stoichiometries such as $CuNiC_4N_4$, NiC_2N_2, $Sb_{2I2}F_{11}$ and $In_3Bi_7(Pb_2S_9)_2$. There is no systematic or obvious trend connecting these misclassified compositions. Examining the receiver operating characteristic (ROC) curve (Figure 1b) and the confusion matrix (Figure 1c) further indicates the ability for the model to successfully differentiate between metals and non-metals. The ROC curve plots the true positive rate (sensitivity), which is the ratio between the number of correctly identify non-metals (true positive) to all the non-metals (the sum of true positive and false negative), against the false positive rate (1–specificity) or the ratio of false positive (incorrectly identified non-metals) to the sum of false positive and true negative (correctly predicted metals). An area under the ROC curve (AUC) of 1.0 represents perfect separation whereas an area of 0.5 (shown by the dashed line) is equivalent to a random guess. Here, the AUC for the classification

of non-metals and metals using our SVC is 0.97 signifying excellent discrimination between these two classes. In comparison, prior methods classifying non-metals and metals using DFT calculated band gaps has reported an AUC of 0.98 (5-fold cross-validated)[27] and 0.93 (test set)[28] suggesting that all models have comparable performance despite using a completely different training sets, descriptors, and machine-learning algorithms. The principle diagonal of the confusion matrix, illustrated in Figure 1c, adds additional support for the model's ability to separate metals from non-metals. It is worth noting that the false positives (FP) are slightly higher than false negatives (FN), 55 compared to 24, meaning that there is a slight preference to mistakenly classify a metal as a non-metal, which is similar to other machine-learning approaches.[27,28] Overall, the confusion matrix illustrates that there is minimal bias and the model has an almost equal success rate for correctly classifying any general composition as a non-metal or metal.

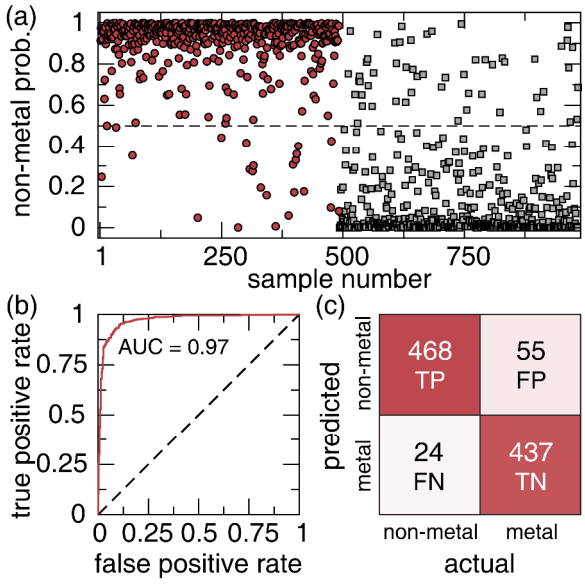


Figure 1. Model performance evaluation for the SVC model based on the test set. (a) Probabilities of a compound to be non-metal shows 92% accuracy. Non-metals and metals are depicted as red circles and grey squares, respectively. (b) Receiver operating characteristic

(ROC) curve for the SVC model with a 0.97 area under the curve (AUC). (c) Confusion matrix illustrating the number of true positives (TP), true negatives (TN), false negatives (FN), and false negatives (FN) with the colors scaled to the number of occurrences.

With the ability to classify any composition as a metal or non-metal using SVC, the E_g for the non-metals was then predicted based on a regression analysis using a ϵ-insensitive support vector regression (SVR)[34,36] method. The objective of ϵ-insensitive SVR is to find a function that contains all the training data within the bound of ±ϵ. Here, among the 3896 experimentally measured, non-zero band gap values (2458 unique compositions) gathered in our database, 80% were again randomly selected as training set, while the remaining 20% were used to test the model. This random selection was performed three times to ensure the statistical validity of our approach. The compositions used for training the regression model range from small band gap semiconductors such as PbSe (E_g = 0.27 eV) to ultra-wide band gap materials like CaF_2 (E_g = 9.92 eV). The resulting SVR achieves excellent agreement between the experimentally measured band gaps and the SVR predicted band gaps for the test set (Figure 2a). The coefficient of determination (r^2) is 0.90 and the root-mean squared error (*RMSE*) is 0.45 eV demonstrating the robust nature of our machine-learning model. The performance of the model (*RMSE*, and r^2) as a function of the number of employed training data as well as the variation of *RMSE* for different ranges of band gaps were also analyzed and are provided in Figure S1. There is a slight underestimation for wide band gap compositions, which most likely is due to a limited number of compounds experimentally reported with these very wide (>6 eV) band gaps. Nevertheless, the performance of our model is similar to the gradient boosting decision tree model based on a property-labelled materials fragment descriptor, which had an r^2 of 0.90 and *RMSE* 0.51 eV[27] and a crystal graph convolutional neural networks using a descriptor set relying on atomic

coordinates derived from the generation of crystal graphs, which resulted in a mean absolute error (*MAE*) of 0.53 eV.[28] The advantage of the approach described here over these statistically comparable models is that our training set is based on experimental band gaps. This provides a significant improvement in the accuracy of predicting E_g with respect to the experimentally measured values compared to machine-learning models trained using DFT calculated band gaps. It is also worth noting that even though our model cannot distinguish between polymorphs due to the absence of structural descriptors, it is quite interesting that all of the machine-learning methods yield similar performance. This further emphasizes the entangled nature of composition and crystal structure. Likewise, the future implementation of structure descriptors in this model will allow researchers to study the influence of composition and structure on fundamental properties such as band gap, which can then be probed independently using these statistical learning models.

Investigating the regression data closer reveals excellent agreement between our machine-leaning model and the experimentally measured band gaps with 589 out of 780 (75%) of all the compositions in our test set are predicted with error of less than 25% and 16% of the compositions have predicted band gap errors of >0.5 eV. The majority of these large errors arise in predicting ultra-wide band gap compositions; one example is the insulator KF which has a predicted band gap of 8.9 eV even though the experimentally reported E_g is 10.3 eV.[37] Again, this discrepancy likely arises from a lack of data for with extremely wide band gaps. Increasing the number of compositions with wide band gaps in the training set should improve the prediction power. There is better agreement for smaller band gap compositions with only a slight overestimation. For example, FeAsS is a narrow band gap semiconductor with a predicted band gap value of 0.18 eV compared to experimental value of 0.06 eV.

One interesting comparison is to relate the machine-learning estimated band gaps to the (PBE-level) DFT calculated band gaps. DFT systematically underestimates the experimental band gap, however, since the SVR model is trained using experimental data, machine learning is expected to also overestimate DFT band gap, similar to the relationship between DFT and experiment.[38,39] As shown in Figure 2b, 11,194 DFT calculated band gaps ($E_{g,DFT}$) available in Materials Project database are plotted against the SVR machine-learning predicted band gaps ($E_{g,SVR}$). As expected, nearly 80% of the $E_{g,SVR}$ values are underestimated compared to the $E_{g,DFT}$, which is consistent with the observed relationship between DFT and the measured band gaps. The 20% of the compounds that fall above the dashed line in Figure 2b indicating $E_{g,DFT} > E_{g,SVR}$ are potentially due to the influence of crystal structure or variations in band gaps of polymorphs. Moreover, analysis of the compositions that fall above the line also indicates that all of the points above the line are ternary and quaternary compositions (all binary compositions fall below the dashed line) suggesting that the descriptor set could be improved to better capture large, complex compositions. In any case, this analysis further supports that our machine-learning model is, in general, much closer to experimentally measured band gaps and does not suffer from the same (DFT) systematic error for a majority of compositions.

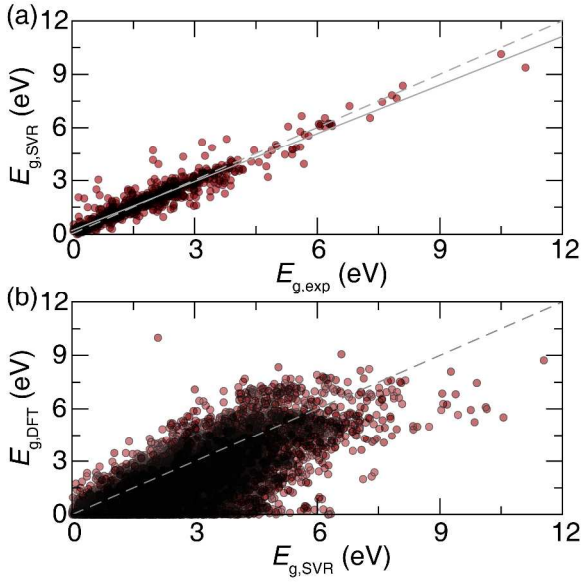


Figure 2. (a) Predicted band gap ($E_{g,SVR}$) versus measured band gap ($E_{g,exp}$) for 780 compounds in test set. The ideal line is shown as dashed line and the fit line is shown as solid line. (b) DFT calculated band gap ($E_{g,DFT}$) versus predicted band gap ($E_{g,SVR}$) for 11,194 compositions.

To quantitatively compare the capability of the SVR machine-learning model, we predicted E_g for 10 specific compounds, unseen by our algorithm, that have also been examined at many different levels of theory including using the PBE functional, the hybrid functional (HSE), the GW (fully self-consistent) method, and a PBE-trained machine learning model reported in AFLOW-ML.[27] These compositions were each selected because they have each been systematically investigated by these levels of theory and are often used to benchmark new exchange and correlation functionals.[14,39,40] Further, the band gaps cover a wide energy range, they span numerous chemical systems and stoichiometries, and many of these compounds are of technological importance. As shown in Table 1, the PBE calculated band gaps ($E_{g,PBE}$) are all substantially underestimated compared to the experiment by an average of ≈60% and a mean average deviation (*MAD*) and *RMSE* of 1.65 eV and 2.1 eV, respectively. The systematic underestimation of DFT with standard exchange and correlation functionals such as LDA or PBE

stems from the well-known discontinuity in the derivative of exchange and correlation functional as the number of electrons deviates from integers.[10,11] Indeed, PBE deviates from the expected straight-line behavior between two integer charges following a convex curvature, referred to as self-interaction. This self-interaction leads to erroneous delocalization and tends to spread the charges across a system. Hartree-Fock, on the other hand, shows a concave behavior for many-electron systems, resulting in an unphysical localization of charges. [41,42] A hybrid exchange and correlation functional, which mixes a fraction of PBE with a fraction Hartree-Fock, produces the almost straight-line behavior, as desired. Although this method does not provide a systematic solution, the band gaps calculated using hybrid functionals ($E_{g,HSE}$) show excellent agreement compared to $E_{g,exp.}$ with *MAD* of 0.67 eV and a *RMSE* 1.0 eV. The agreement between computation and experiment is undoubtedly improved; however, using hybrid functionals increases the computational cost by several orders of magnitude compared to PBE. Another common, beyond DFT approach to solve the band gap problem is considering quasi-particle energies to solve the self-interaction energy through perturbation theory. The *GW*-type calculated band gaps ($E_{g,GW}$) are the most involved, and computationally expensive of the methods examined although they provide an extremely accurate estimate of the experimental band gap with a *MAD* of only 0.33 eV and a *RMSE* 0.33 eV, as shown in Table 1. Comparing these first principles calculated results with the band gaps predicted by the AFLOW machine-learning algorithm ($E_{g,AFLOW}$) shows AFLOW-ML also significantly underestimates the $E_{g,exp}$ with *MAD* of only 1.60 eV and a *RMSE* of 2.25 eV, which is expected given that $E_{g,AFLOW}$ is trained using the $E_{g,PBE}$ values.

Table 1. Comparison of experimentally measured band gap values, with the band gaps calculated using the PBE functional, a hybrid functional (HSE), the *GW* approach, the AFLOW-ML project, and the SVR model presented here. The percentage for each calculated value is with respect to the experimentally measured E_g. The mean absolute deviation (*MAD*) and root mean square error (*RMSE*) are also shown compared to experimental band gap. The band gap units are all in eV.

composition	$E_{g,exp.}$	$E_{g,PBE}$	$E_{g,HSE}$	$E_{g,GW}$	$E_{g,AFLOW}$	$E_{g,SVR}$
GaN	3.2[43]	1.62 (−49%)[43]	3.14 (−2%)[16]	3.32 (4%)[43]	1.85 (−42%)	4.45 (39%)
CdTe	1.6[44]	0.62 (−61%)[45]	1.52 (−3%)[45]	1.76 (12%)[44]	0.67 (−57%)	1.43 (−9%)
LiF	14.2[43]	9.2 (−35%)[43]	11.47 (−19%)[46]	15.1 (6%)[43]	8.27 (−42%)	9.87 (−30%)
TiO_2	3.42[14]	2.13 (−37%)[14]	3.67 (7%)[14]	3.73 (9%)[14]	2.09 (−38%)	3.99 (16%)
$CuSbS_2$	1.38[47]	0.9 (−35%)[47]	1.69 (22%)[47]	1.1 (−20%)[47]	0.79 (−42%)	1.39 (1%)
ZnS	3.91[43]	2.07 (−47%)[43]	3.49 (−11%)[16]	4.15 (6%)[43]	2.6 (−33%)	3.12 (−20%)
Cu_2ZnSnS_4	1.6[48]	0.28 (−83%)[48]	0.09 (−94%)[48]	1.64 (3%)[48]	N/A	1.75 (9%)
PbTe	0.19[43]	0 (−100%)[43]	0.19 (0%)[49]	0.26 (36%)[43]	0 (−100%)	0.2 (5%)
GaAs	1.52[43]	0.19 (−86%)[43]	1.12 (−26%)[16]	1.52 (0%)[43]	0.24 (−84%)	1.28 (−15%)
ZnO	3.44[43]	0.67 (−81%)[43]	2.49 (−28%)[16]	3.2 (−7%)[43]	1.87 (−46%)	3.41 (−1%)
MAD		1.65	0.67	0.22	1.60	0.75
RMSE		2.1	1.0	0.33	2.25	1.46

Analyzing the band gaps using the new SVR model ($E_{g,SVR}$) shows remarkable agreement with $E_{g,exp}$. The *MAD* is 0.75 eV and the *RMSE* is 1.46 eV indicate the error of this model falls approximately halfway between PBE functional and the hybrid functional. Among the selected compounds, our method performs admirably for compounds such as $CuSbS_2$ (1% error), PbTe (5% error), and ZnO (1% error). The only place the machine-learning method appears to perform

poorly is in predicting the ultra-wide band gap of LiF. However, the error in this case is still only 30% with a predicted value of 9.87 eV but a $E_{g,exp}$ of 14.2 eV. The deviation for LiF is not surprising considering our method tends to underestimate high band gap values, which can be partially attributed to the imbalance of input data with significantly fewer data points for high band gap values. More importantly, comparing $E_{g,SVR}$ of the 10 compounds shows that our machine-learning model performs significantly better than DFT calculations employing the PBE functional and AFLOW-ML. It is possible that improving the prediction of the ultra-wide band gap materials will push this SVR approach to the level of the hybrid functional or beyond.

In summary, this contribution presents a new supervised machine-learning scheme trained using 3896 experimentally measured band gaps extracted from the chemical literature. Our results show that this method, using a descriptor set based on only composition information, is capable of discriminating metals from non-metals with outstanding accuracy (≈92%). The band gap of the compounds classified as non-metals were then predicted using a support vector regression approach. The accuracy of this model outperforms standard DFT calculations. These results are significant because they present a SVR-based model that can reliably predict the band gap of any material at a drastically reduced computational cost compared to higher levels of theory. Finally, the success of this method enables us to employ our model to estimate the band gap of 94,095 compounds compiled in PCD. We first classify the compound as a metal or non-metal, followed by a quantitative prediction of the band gap values for the non-metals. This large dataset is available as Supporting Information for researchers to use in their development of new inorganic materials.

METHODS: The machine-learning model was created using Support Vector classification (SVC) and Support Vector regression (SVR) methods. These methods were carried out using using PLS_Toolbox Version 8.5 (Eigenvector Research Inc., Wenatchee, WA) within the MATLAB® (R2017a) computational environment. Both SVC[33,34] and SVR[34,36] implemented a RBF as the kernel function using a 10-fold venetian blinds cross-validation method, with normalization and autoscaling applied to training set prior to learning. For SVC, cost constant (C) and free parameter (γ) were optimized to 32 and 0.01. Similarly, C and γ were optimized to 10 and 0.01 for SVR respectively while ϵ was set at 0.1. The descriptors were developed based on only compositional data and their mathematical expressions as described in the text. The experimental band gap values were extracted from literature[29–32] as referenced and for metals they were extracted from Materials Project database[35].

AUTHOR INFORMATION

Corresponding Author

*E-mail: jbrgoch@uh.edu

ORCID

Ya Zhuo: 0000-0003-2554-498X

Aria Mansouri Tehrani: 0000-0003-1968-0379

Jakoah Brgoch: 0000-0002-1406-1352

Notes

The authors declare no competing financial interests.

ACKNOWLEDGMENT

The authors thank the Department of Chemistry and the Division of Research at the University of Houston for providing generous start-up funds as well as the National Science Foundation through No. NSF-CMMI 15-62142 for supporting this research. The authors also thank Dr. Anton Oliynyk for many fruitful discussions.

ASSOCIATED CONTENT

Supporting Information Available:

Complete set of the employed variables (Table S1), a comparison of different machine-learning methods (Table S2), and the *RMSE* and r^2 as a function of training set size and the *RMSE* for different band gap windows (Figure S1).

The complete training set of experimental band gaps

The predicted band gaps of compounds from Pearson's Crystal Database (PCD)

REFERENCES

(1) Fasol, G. Room-Temperature Blue Gallium Nitride Laser Diode. *Science* **1996**, *272*, 1751–1752.

(2) Schubert, E. F.; Kim, J. K. Solid-State Light Sources Getting Smart. *Science* **2005**, *308*, 1274–1278.

(3) Radisavljevic, B.; Radenovic, A.; Brivio, J.; Giacometti, V.; Kis, A. Single-Layer MoS_2 Transistors. *Nat. Nanotechnol.* **2011**, *6*, 147–150.

(4) Polman, A.; Knight, M.; Garnett, E. C.; Ehrler, B.; Sinke, W. C. Photovoltaic Materials: Present Efficiencies and Future Challenges. *Science* **2016**, *352*, aad4424.

(5) Ahn, S.; Jung, S.; Gwak, J.; Cho, A.; Shin, K.; Yoon, K.; Park, D.; Cheong, H.; Yun, J. H. Determination of Band Gap Energy (Eg) of $Cu_2ZnSnSe_4$ Thin Films: On the Discrepancies of Reported Band Gap Values. *Appl. Phys. Lett.* **2010**, *97*, 21905.

(6) Canning, A.; Chaudhry, A.; Boutchko, R.; Grønbech-Jensen, N. First-Principles Study of Luminescence in Ce-Doped Inorganic Scintillators. *Phys. Rev. B* **2011**, *83*, 125115.

(7) Ceder, G.; Chiang, Y. M.; Sadoway, D. R.; Aydinol, M. K.; Jang, Y. I.; Huang, B. Identification of Cathode Materials for Lithium Batteries Guided by First-Principles Calculations. *Nature* **1998**, *392*, 694–696.

(8) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The High-Throughput Highway to Computational Materials Design. *Nat. Mater.* **2013**, *12*, 191–201.

(9) Setyawan, W.; Gaume, R. M.; Lam, S.; Feigelson, R. S.; Curtarolo, S. High-Throughput Combinatorial Database of Electronic Band Structures for Inorganic Scintillator Materials. *ACS Comb. Sci.* **2011**, *13*, 382–390.

(10) Perdew, J. P. Density Functional Theory and the Band Gap Problem. *Int. J. Quantum Chem.* **2009**, *28*, 497–523.

(11) Seidl, A.; Görling, A.; Vogl, P.; Majewski, J. A.; Levy, M. Generalized Kohn-Sham Schemes and the Band-Gap Problem. *Phys. Rev. B* **1996**, *53*, 3764–3774.

(12) Heyd, J.; Scuseria, G. E. Efficient Hybrid Density Functional Calculations in Solids: Assessment of the Heyd–Scuseria–Ernzerhof Screened Coulomb Hybrid Functional. *J. Chem. Phys.* **2004**, *121*, 1187–1192.

(13) Garza, A. J.; Scuseria, G. E. Predicting Band Gaps with Hybrid Density Functionals. *J.*

Phys. Chem. Lett. **2016**, *7*, 4165–4170.

(14) Gerosa, M.; Bottani, C. E.; Caramella, L.; Onida, G.; Di Valentin, C.; Pacchioni, G. Electronic Structure and Phase Stability of Oxide Semiconductors: Performance of Dielectric-Dependent Hybrid Functional DFT, Benchmarked against GW Band Structure Calculations and Experiments. *Phys. Rev. B* **2015**, *91*, 155201.

(15) Becke, A. D.; Johnson, E. R. A Simple Effective Potential for Exchange. *J. Chem. Phys.* **2006**, *124*, 221101.

(16) Tran, F.; Blaha, P. Accurate Band Gaps of Semiconductors and Insulators with a Semilocal Exchange-Correlation Potential. *Phys. Rev. Lett.* **2009**, *102*, 226401.

(17) Chan, M. K. Y.; Ceder, G. Efficient Band Gap Prediction for Solids. *Phys. Rev. Lett.* **2010**, *105*, 196403.

(18) Koller, D.; Tran, F.; Blaha, P. Merits and Limits of the Modified Becke-Johnson Exchange Potential. *Phys. Rev. B* **2011**, *83*, 195134.

(19) Pilania, G.; Gubernatis, J. E.; Lookman, T. Multi-Fidelity Machine Learning Models for Accurate Bandgap Predictions of Solids. *Comput. Mater. Sci.* **2017**, *129*, 156–163.

(20) Himmetoglu, B.; Floris, A.; Gironcoli, S.; Cococcioni, M. Hubbard-Corrected DFT Energy Functionals: The LDA+U Description of Correlated Systems. *Int. J. Quantum Chem.* **2014**, *114*, 14–49.

(21) Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B. P.; Ramprasad, R.; Gubernatis, J. E.; Lookman, T. Machine Learning Bandgaps of Double Perovskites. *Sci. Rep.* **2016**, *6*, 19375.

(22) Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; Tanaka, I. Prediction Model of Band Gap for Inorganic Compounds by Combination of Density Functional Theory Calculations and

Machine Learning Techniques. *Phys. Rev. B* **2016**, *93*, 115104.

(23) Dey, P.; Bible, J.; Datta, S.; Broderick, S.; Jasinski, J.; Sunkara, M.; Menon, M.; Rajan, K. Informatics-Aided Bandgap Engineering for Solar Materials. *Comput. Mater. Sci.* **2014**, *83*, 185–195.

(24) Zhang, Z.; Peng, R.; Chen, N. Artificial Neural Network Prediction of the Band Gap and Melting Point of Binary and Ternary Compound Semiconductors. *Mater. Sci. Eng. B* **1998**, *54*, 149–152.

(25) Gu, T.; Lu, W.; Bao, X.; Chen, N. Using Support Vector Regression for the Prediction of the Band Gap and Melting Point of Binary and Ternary Compound Semiconductors. *Solid State Sci.* **2006**, *8*, 129–136.

(26) Curtarolo, S.; Setyawan, W.; Hart, G. L. W.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; et al. AFLOW: An Automatic Framework for High-Throughput Materials Discovery. *Comput. Mater. Sci.* **2012**, *58*, 218–226.

(27) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nat. Commun.* **2017**, *8*, 15679.

(28) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for Accurate and Interpretable Prediction of Material Properties. *arXiv* **2017**, *1710*.

(29) Kiselyova, N. N.; Dudarev, V. A.; Korzhuyev, M. A. Database on the Bandgap of Inorganic Substances and Materials. *Inorg. Mater. Appl. Res.* **2016**, *7*, 34–39.

(30) Strehlow, W. H.; Cook, E. L. Compilation of Energy Band Gaps in Elemental and Binary Compound Semiconductors and Insulators. *J. Phys. Chem. Ref. Data* **1973**, *2*, 163–200.

(31) Joshi, N. V. *Photoconductivity : Art, Science, and Technology*; Marcel Dekker: New York,

1990.

(32) Madelung, O. *Semiconductors : Data Handbook*; Springer: New York, 2004.

(33) Cortes, C.; Vladimir, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.

(34) Chang, C. C.; Lin, C. J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27.

(35) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, 11002.

(36) Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A. J.; Vapnik, V. Support Vector Regression Machines. *Adv. Neural Inf. Process. Syst.* **1996**, *1*, 155–161.

(37) Poole, R. T.; Jenkin, J. G.; Liesegang, J.; Leckey, R. C. G. Electronic Band Structure of the Alkali Halides. I. Experimental Parameters. *Phys. Rev. B* **1975**, *11*, 5179–5189.

(38) Shishkin, M.; Marsman, M.; Kresse, G. Accurate Quasiparticle Spectra from Self-Consistent GW Calculations with Vertex Corrections. *Phys. Rev. Lett.* **2007**, *99*, 246403.

(39) Crowley, J. M.; Tahir-Kheli, J.; Goddard, W. A. Resolution of the Band Gap Prediction Problem for Materials Design. *J. Phys. Chem. Lett.* **2016**, *7*, 1198–1203.

(40) Clark, S. J.; Robertson, J. Screened Exchange Density Functional Applied to Solids. *Phys. Rev. B* **2010**, *82*, 85208.

(41) Cohen, A. J.; Mori-Sanchez, P.; Yang, W. Insights into Current Limitations of Density Functional Theory. *Science* **2008**, *321*, 792–794.

(42) Mori-Sanchez, P.; Cohen, A. J.; Yang, W. Localization and Delocalization Errors in Density Functional Theory and Implications for Band-Gap Prediction. *Phys. Rev. Lett.* **2008**, *100*, 146401.

(43) Shishkin, M.; Kresse, G. Self-Consistent GW Calculations for Semiconductors and Insulators. *Phys. Rev. B* **2007**, *75*, 235102.

(44) Zakharov, O.; Rubio, A.; Blase, X.; Cohen, M. L.; Louie, S. G. Quasiparticle Band Structures of Six II-VI Compounds: ZnS, ZnSe, ZnTe, CdS, CdSe, and CdTe. *Phys. Rev. B* **1994**, *50*, 10780–10787.

(45) Heyd, J.; Peralta, J. E.; Scuseria, G. E.; Martin, R. L. Energy Band Gaps and Lattice Parameters Evaluated with the Heyd-Scuseria-Ernzerhof Screened Hybrid Functional Energy Band Gaps and Lattice Parameters Evaluated with the Heyd- Scuseria-Ernzerhof Screened Hybrid Functional. *J. Chem. Phys.* **2005**, *123*, 174101.

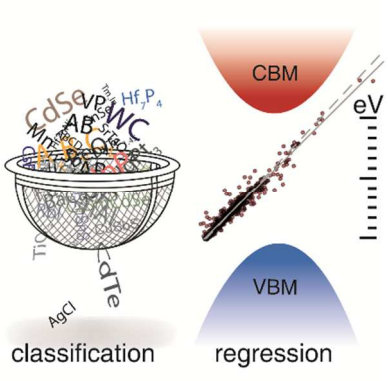
(46) Schimka, L.; Harl, J.; Kresse, G. Improved Hybrid Functional for Solids: The HSEsol Functional. *J. Chem. Phys.* **2011**, *134*, 24116.

(47) Yu, L.; Kokenyesi, R. S.; Keszler, D. A.; Zunger, A. Inverse Design of High Absorption Thin-Film Photovoltaic Materials. *Adv. Energy Mater.* **2013**, *3*, 43–48.

(48) Zhao, H.; Persson, C. Optical Properties of Cu(In,Ga)Se_2 and $Cu_2ZnSn(S,Se)_4$. *Thin Solid Films* **2011**, *519*, 7508–7512.

(49) Skelton, J. M.; Parker, S. C.; Togo, A.; Tanaka, I.; Walsh, A. Thermal Physics of the Lead Chalcogenides PbS, PbSe, and PbTe from First Principles. *Phys. Rev. B* **2014**, *89*, 205203.

Table of contents figure



SUPPORTING INFORMATION

Predicting the Band Gaps of Inorganic Solids by Machine Learning

*Ya Zhuo, Aria Mansouri Tehrani, Jakoah Brgoch**
Department of Chemistry, University of Houston, Houston, Texas 77204, United States

Table S1. Variables used in the machine learning model. Each compositional property is divided into 4 variables according to sum, difference, largest, and smallest values of the composition's constituent elements. These values were extracted from open sources and databases.[1–12]

All 136 used variables.
1–4. Atomic number
5–8. Atomic weight
9–12. Period number
13–16. Group number
17–20. Family number
21–24. L quantum number
25–28. Mendeleev number
29–32. Atomic radius (Å)
33-36. Covalent radius (Å)
37–40. Zunger radius (Å)
41-44. Ionic radius (Å)
45-48. Crystal radius (Å)
49-52. Pauling EN
53-56. Martynov-Batsanov EN
57-60. Gordy EN
61-64. Mulliken EN
65-68. Allen EN
69-72. Metallic valence
73-76. Number of valence electrons
77-80. Number of s electrons
81-84. Number of p electrons
85-88. Number of d electrons
89-92. Number of outer shell electrons
93-96. First ionization energy (kJ/mol)
97-100. Polarizability
101-104. Melting point (K)
105-108. Boiling point (K)
109-112. Density (g/mL)
113-116. Specific heat (J/g•K)
117-120. Heat of fusion (kJ/mol)
121-124. Heat of vaporization (kJ/mol)
125-128. Thermal conductivity (W/m•K)
129-132. Heat atomization (kJ/mol)
133-136. Cohesive energy (eV)

Table S2. The performance of different machine learning methods by using identical training data and test data. The regressors include SVR with a radial basis function kernel (SVR_{rbf}), with a linear kernel (SVR_{lin}), and with a polynomial kernel (SVR_{pol}), K-Nearest Neighbors (KNN), and kernel ridge regression (KRR). The classifiers include SVC with a radial basis function kernel (SVC_{rbf}), with a linear kernel (SVC_{lin}), and with a polynomial kernel (SVC_{pol}), K-Nearest Neighbors (KNN), and logistic regression (LR).

regressor	RMSE (eV)	r^2	classifier	accuracy
SVR_{rbf}	0.45	0.90	SVC_{rbf}	0.92
SVR_{lin}	0.73	0.72	SVC_{lin}	0.85
SVR_{pol}	0.46	0.89	SVC_{pol}	0.90
KNN	0.54	0.85	KNN	0.88
KRR	0.72	0.74	LR	0.86

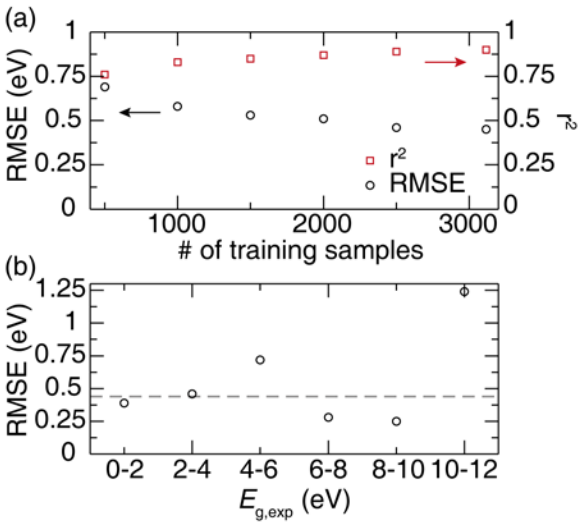


Figure S1. (a) Root mean square error (*RMSE*) (black circles) and coeeficient of determination (r^2) (red squares) of models that are trained by different number of samples in training set. (b) *RMSE* and r^2 vary across the bandgap range. The average *RMSE* is shown as dashed line.

REFERENCES

(1) Zunger, A. Systematization of the Stable Crystal Structure of All AB-Type Binary Compounds: A Pseudopotential Orbital-Radii Approach. *Phys. Rev. B* **1980**, *22*, 5839–5872.

(2) Slater, J. C. Atomic Radii in Crystals. *J. Chem. Phys.* **1964**, *411*, 1300–2686.

(3) Sanderson, R. T. *Chemical Periodicity*; Reinhold Pub. Corp.: New York, **1960**.

(4) Porterfield, W. W. Inorganic Chemistry, a Unified Approach. In *Inorganic Chemistry, a Unified Approach*; Academic Press: San Diego, **1993**.

(5) Pauling, L. The Nature of the Chemical Bond. IV. The Energy of Single Bonds and the Relative Electronegativity of Atoms. *J. Am. Chem. Soc.* **1932**, *54*, 3570–3582.

(6) Mulliken, R. S. A New Electroaffinity Scale; Together with Data on Valence States and on Valence Ionization Potentials and Electron Affinities. *J. Chem. Phys.* **1934**, *2*, 1833–767.

(7) Martynov, A. I.; Batsanov, S. S. A New Approach to the Determination of the Electronegativity of Atoms. *Russ. J. Inorg. Chem.* **1980**, *25*, 1737–1739.

(8) James, A. M.; Lord, M. P. *Macmillan's Chemical and Physical Data*; Macmillan: Basingstoke, **1992**.

(9) Ghosh, D. C.; Chakraborty, T. Gordy's Electrostatic Scale of Electronegativity Revisited. *J. Mol. Struct. THEOCHEM* **2009**, *906*, 87–93.

(10) Kaye, G. W. C.; Laby, T. H. *Tables of Physical and Chemical Constants and Some Mathematical Functions*; Longmans, Green and Company: London, **1921**.

(11) Ellis, H. *Nuffield Advanced Science: Book of Data*; Longman: London, **1972**.

(12) Allen, L. C. Electronegativity Is the Average One-Electron Energy of the Valence-Shell Electrons in Ground-State Free Atoms. *J. Am. Chem. Soc* **1989**, *1*, 9003–9014.