Machine learning the band gap properties of kesterite I_2-II-IV-V_4 quaternary compounds for photovoltaics applications

L. Weston[1,2] and C. Stampfl[1]

[1]*School of Physics, The University of Sydney, Sydney, New South Wales 2006, Australia*

[2]*Materials Department, University of California, Santa Barbara, California 93106-5050, USA*

(Received 29 August 2017; revised manuscript received 16 April 2018; published 21 August 2018)

Kesterite I_2-II-IV-V_4 semiconductors are promising solar absorbers for photovoltaics applications. The band gap and its character, either direct or indirect, are fundamental properties determining photovoltaic-device efficiency. We use a combination of accurate first-principles calculations and machine learning to predict the properties of the band gap for a large number of kesterite I_2-II-IV-V_4 semiconductors. In determining the magnitude of the fundamental gap, we compare results for a number of machine-learning models, and achieve a root mean squared error as low as 283 meV; the best results are achieved using support-vector regression with a radial-bias kernel. This error is well within the uncertainty of even the most advanced first-principles methods for calculating semiconductor band gaps. Predicting the direct-indirect property of the band gap is more challenging. After significant feature engineering, we are able to train a classifier that predicts the nature of the band gap with an accuracy of 89% using logistic regression. Using these trained models, the band gap properties of 1568 kesterite I_2-II-IV-V_4 compounds are predicted. We find 717 compounds with band gaps in the range 0.5–2.5 eV that can potentially act as solar absorbers and 242 materials with a band gap in the "*optimum range*" of 1.2–1.8 eV. The stability of these 242 compounds is assessed by calculating the energy above hull using the Materials Project database, and the band gaps are verified using hybrid functional calculations; in the end, we identify 25 compounds that are expected to be synthesizable and have a band gap in the range 1.2–1.8 eV—most of which are previously unexplored. These results will be useful in the materials engineering of efficient photovoltaic devices.

DOI: 10.1103/PhysRevMaterials.2.085407

I. INTRODUCTION

Quaternary I_2-II-IV-VI_4 semiconductors offer a unique opportunity in materials engineering due to the vast design space [1,2]. The different cation-anion combinations exhibit band gaps spanning the visible spectrum. This tunability in the band gap upon cation and anion mutation has led to intense interest in these materials for applications as solar absorbers for photovoltaic devices [3–5].

For photovoltaics applications, the band gap is a fundamental property determining efficiency, with band gaps around 1.5 eV being the most efficient solar absorbers [6]. Moreover, the direct-indirect character of the band gap is of a fundamental importance: while direct gap materials are typically stronger absorbers than indirect materials, they may also have shorter photocarrier lifetimes and suffer from carrier recombination [7]. Given the complex design space of I_2-II-IV-VI_4 compounds, it becomes difficult to characterize all of the possible cation-anion combinations, both theoretically and experimentally. There exists multiple possibilities for both the cation ordering, including kesterite and stannite, as well as the crystal symmetry, since the geometry may be derived from either the zinc blende or wurtzite phase [8]; consequently, there are thousands if not tens of thousands of possible I_2-II-IV-VI_4 materials.

From a theoretical perspective, the calculation of semiconductor band gaps within traditional density functional theory (DFT) suffers from the well-known underestimation error [9]. This can be overcome with more accurate theoretical approaches such as screened hybrid functionals [10] or many-body perturbation theory (GW) [11]. However, these are far more computationally expensive, and therefore are difficult to implement on a large set of materials. Indeed, the available large databases of semiconductor band gaps mostly rely on traditional DFT calculations within the generalized-gradient approximation (GGA) [12–14].

One possible approach to overcome this challenge is to use machine learning to generate or improve predictions [15,16]. By performing accurate high-level first-principles calculations on a subset of I_2-II-IV-VI_4 compounds, the results can be used to train a machine-learning model to predict the properties of the remaining materials in the design space. Lee *et al.* used machine learning to predict the band gaps of 156 AX binary compounds using element-specific descriptors including the band gap from low-level DFT calculations, achieving a root mean squared error (RMSE) of 180 meV with support-vector regression [17]. Pilania *et al.* used kernel-ridge regression to predict the band gaps of 1306 double perovskites and achieved a RMSE of 80 meV using a 16-dimensional set of element-specific descriptors [18]. Ward *et al.* proposed a large set of 140 universal descriptors to predict band gaps from a very large data set, and identified new possible solar absorbers [19]; it was also found that model accuracy was improved by partitioning the data set into groups of similar materials, suggesting that machine-learning predictions would work best on isostructural and isoelectronic materials. To the best of our knowledge, classification of band gaps as either direct or indirect has not been attempted from a machine-learning perspective.

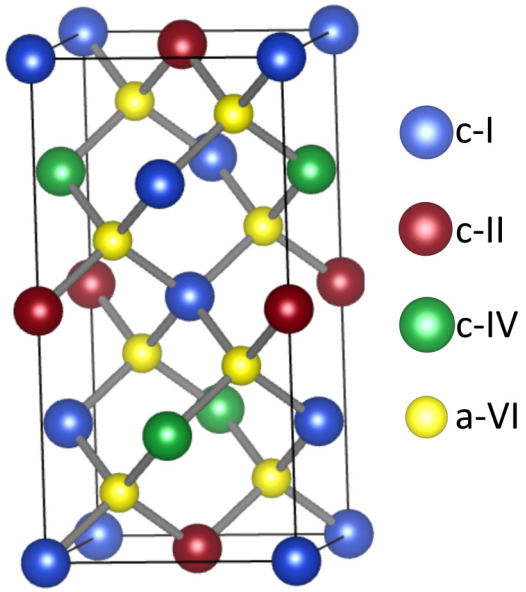

FIG. 1. Zinc-blende-based kesterite structure for a I_2-II-IV-VI_4 compound. The cations with valence I (c-I), II (c-II), and IV (c-IV) are indicated by blue, red, and green spheres, respectively. The anion has a valence VI (a-VI) and is indicated by a yellow sphere.

In the present paper, we study I_2-II-IV-VI_4 semiconductors in the zinc-blende-based kesterite structure. This provides the opportunity to study a large number of materials systems that are both isostructural and isoelectronic. We consider 1568 possible cation-anion combinations, and perform accurate hybrid functional calculations for the band gaps on a randomly selected subset of 200 materials; these results are then used to train various machine-learning models. Using support-vector regression with a radial bias kernel, we are able to predict the magnitude of the fundamental gap with a RMSE of 283 meV, using only three simple, element-specific descriptors per element in the compound. We find that classification of these materials as direct or indirect semiconductors is more challenging. After substantial feature engineering, we train a classifier with an accuracy of 89% using logistic regression. The trained models are used to predict the band gap properties for all 1568 compounds, and to identify potential solar absorbers; these results will be useful in the design and engineering of kesterite I_2-II-IV-VI_4 semiconductors for photovoltaics applications.

II. METHODOLOGY

A. Materials systems

The kesterite structure is derived via cation mutation of the II-VI binary zinc blende phase, and is shown in Fig. 1. For the I_2-II-IV-VI_4 compounds, we consider the following: I = Li, Na, K, Rb, Cs, Cu, Ag; II = Be, Mg, Ca, Sr, Ba, Zn, Cd, Hg; IV = C, Si, Ge, Sn, Ti, Zr, Hf; VI = O, S, Se, Te. This provides a total of 1568 compounds. While a number of these compounds will not be thermodynamically stable, they are still useful in training the machine-learning models. We randomly select a subset of 200 materials, and calculate their band gap properties.

B. First-principles calculations

Our calculations are performed in the Vienna *Ab initio* Simulation Package (VASP) [20], using density functional theory (DFT) within the generalized Kohn-Sham scheme [21]. The valence electrons are separated from the core by use of projector augmented wave (PAW) potentials [22].

The lattice parameters and the internal ionic coordinates are determined by a full relaxation of the cell using the PBEsol functional [23]; PBEsol has been shown to give highly accurate geometries for zinc blende semiconductors [24]. Once the geometry has been determined, we perform a fixed-point calculation of the band gap using the screened hybrid functional of Heyd, Scuseria, and Ernzerhof (HSE) [25,26]. In this approach, the short-range exchange potential is calculated by mixing a fraction of nonlocal Hartree-Fock exchange with the GGA of Perdew, Burke, and Ernzerhof (PBE) [27]. The long-range exchange potential and the correlation potential are calculated with PBE. The screening parameter is set to 0.2 Å^{-1} and the mixing parameter to $\alpha = 0.25$. The HSE functional provides highly accurate semiconductor band gaps when compared to traditional DFT [10]. For the stability analysis in Sec. III C, we use the PBE functional in the formation enthalpy calculations to be more consistent with the Materials Project database [13].

The kesterite phase has an eight-atom body-centered tetragonal (BCT) primitive cell. For the geometry relaxation within PBEsol, an $8 \times 8 \times 8$ Monkhorst-Pack k-point grid is used for integrations over the Brillouin zone [28]. For determination of the band gap within HSE, we perform a full calculation along the high-symmetry path in the BCT Brillouin zone [29]. We use a plane wave cutoff of 400 eV for the sulfides, selenides, and tellurides, and 500 eV for the oxides. For the selenides and tellurides, the spin-orbit splitting (Δ_{SO}) at the valence band maximum is neglected; however, the splitting only affects the band gap by $\Delta_{SO}/3$, and therefore we expect an error less than 100 meV in most cases [30].

C. Machine-learning models

1. Regression

The magnitude of the band gap can be predicted using a number of regression models. Regression aims to determine a relationship between the features of each compound, called descriptors (discussed below), and the band gap of the material. We present the key feature of each model below.

Linear and support-vector regression. Consider a linear function $y = \langle \omega, x \rangle + b$, where ω and x are vectors and $\langle ., . \rangle$ denotes a dot product; for a set of features x_i and outcomes y_i, an ordinary least squares regression will attempt to fit ω and b to minimize the sum of squares $\sum_i [y_i - (\langle \omega, x_i \rangle + b)]^2$. Support-vector regression introduces the concept of a margin ε, and attempts to fit a curve such that all of the points lie within the margin. Support-vector regression also favors curve "*flatness*," by reducing the sensitivity to outliers. The problem of support-vector regression is typically written in the following way [31]:

$$\begin{aligned} &\text{minimize} : \ ||\omega^2||, \\ &\text{subject to} : \ |y_i - (\langle \omega, x_i \rangle + b)| \leqslant \varepsilon. \end{aligned} \tag{1}$$

Here, minimizing $||\omega^2||$ maximizes the *flatness* in the curve. In many cases, it is not possible to fit a curve such that $|y_i - (\langle \omega, x_i \rangle + b)| \leqslant \varepsilon$, and therefore additional parameters are introduced to construct a so-called "*soft margin*." There is a trade-off between the *softness* in the margin and the *flatness* in the curve that is determined by a constant known as the C parameter, and this must be tuned to optimize predictions.

In addition to linear support-vector machines, a nonlinear transformation may be applied on the feature space by the so-called "*kernel trick*." We implement support-vector regression with a radial bias function. For two data points x and x', this function is defined as follows:

$$R(x, x') = \exp(-\gamma ||x - x'||). \tag{2}$$

The parameter γ determines how quickly R decays with the distance between x and x' in the feature space; γ can be tuned to optimize predictions.

Tree-based methods. The most simple tree-based regression is a decision tree regressor [32]. For a set of inputs x_i and outcomes y_i, the decision tree will split regions in the feature space into two groups, R_1 and R_2, having mean outcomes $\widehat{y_1}$ and $\widehat{y_2}$, in such a way that minimizes the residual sum squared (RSS),

$$\text{RSS} = \sum_{i \in R_1} (y_i - \widehat{y_1})^2 + \sum_{i \in R_2} (y_i - \widehat{y_2})^2. \tag{3}$$

After the initial split, the tree will continue to make further optimum splits in the feature space until some convergence criteria is met, and the remaining unsplit groups R_t are called terminal nodes, or leaves. The total function to minimize for the decision tree regressor (DTR) is the following,

$$\text{DTR} = \sum_{t=1}^{T} \sum_{i \in R_t} (y_i - \widehat{y_t})^2 + \gamma T, \tag{4}$$

where T is the number of terminal nodes and $\widehat{y_t}$ is the mean of all y_i in R_t. The second term γT is a penalty that prevents overfitting for more complex trees.

Decision trees are computationally efficient and easy to interpret, however often suffer from over fitting and inaccurate predictions. Ensemble-tree based methods are typically used to overcome the shortcomings of simple decision trees. One such method is the random forest regressor [33]. Random forest regression works using the technique of bootstrap aggregating, in which a random subset of x_i and y_i are chosen to train a decision tree; this process is repeated to fit many trees, and then predictions can be made by averaging the results from the ensemble of regression trees. The number of randomly selected decision trees to be fitted is a parameter that is typically tuned to optimize predictions. Another ensemble method is the gradient-boosted regression tree [34]. Boosting is a technique in which many individual decision trees are trained sequentially; each tree is trained from the residuals of the previous tree, as defined by Eq. (4). In this way, the new tree that is added to the ensemble is the one that best minimizes the residuals. Ensemble methods such as random forest and boosting typically correct for overfitting, and reduce the sensitivity to noise in the training set.

2. Classification

We train a binary direct-indirect band gap classifier using logistic regression. In this approach, the model can predict the probability of a binary outcome, and makes predictions on the outcome by determining which is more likely. The logistic function $L(x)$ is an S-shaped curved that varies smoothly between zero and 1; $L(x)$ takes the vector of features x, and if $L(x) < 0.5$, the classifier will predict a binary outcome of zero; when $L(x) > 0.5$, the classifier predicts a binary outcome of 1. Similar to linear regression, which attempts to fit the optimum coefficients for the linear equation $\langle \omega, x \rangle + b$, logistic regression is fit by optimizing the coefficients in $L(x)$,

$$L(x) = \frac{1}{1 + \exp[-(\langle \omega, x \rangle + b)]}, \tag{5}$$

by minimizing the number of incorrect classifications on the training data. In the present study, the binary outcomes are direct-indirect rather than 0-1.

3. Feature space

A number of different features have been proposed as predictors for materials properties [17–19,35]. In the present work, we first use a simple set of element-specific features. For each of the elements in the I_2-II-IV-VI_4 compound, the electronegativity, ionic radius, and row in the Periodic Table are used; this gives 12 features total per compound. This 12-dimensional feature space works extremely well for predicting the magnitude of the band gap using regression techniques. However, this set of features performed poorly when implemented in the direct-indirect band gap classifier, and we had to perform substantial feature engineering, as will be discussed in Sec. III B.

III. RESULTS AND DISCUSSION

Of the 200 compounds studied, 16 either did not have a band gap, or did not converge at some stage of the calculation; these were excluded from the fitting. The band gaps of the remaining 184 I_2-II-IV-VI_4 compounds are used to train the machine-learning models. We first discuss determination of the magnitude of the fundamental gap using regression models. Next, we will discuss training of a classifier to determine the direct-indirect character of the gap.

A. Band gap regressor

A number of regression models are used to fit the magnitude of the band gap. Where appropriate we performed feature normalization, and performed a search over any tunable parameters to optimize the regressor. The accuracy of the model is determined using 10-fold cross validation. The accuracy of the model is assessed by analyzing the root mean squared error (RMSE), and the R^2 coefficient of determination. The results for each regression model are presented in Table I.

Linear regression, which is the simplest model considered, gave a RMSE of 0.59 eV. This error is larger than desired. Support-vector regression with a linear kernel gave almost the same error. However, upon training a support-vector machine

TABLE I. Root mean squared error (RMSE) and R^2 value for machine-learning models based on 10-fold cross validation. Results are shown for linear regression, support-vector regression with a linear (SVR-L) and radial bias function (SVR-R) kernel, decision tree, random forest, and boosted regression tree (Boosted reg. tree).

Model	R^2	RMSE (eV)
Linear regression	0.796	0.590
SVR-L	0.789	0.592
SVR-RBF	0.957	0.283
Decision tree	0.823	0.492
Random forest	0.874	0.435
Boosted reg. tree	0.934	0.358

with a radial bias kernel, this error is greatly reduced; we find a RMSE of only 0.283 meV, and R^2 of 0.957, suggesting an excellent fit.

For the regression-tree-based methods, as expected, the simple decision tree gives the largest RMSE; the RMSE is reduced for the random forest regressor. The boosted regression tree gave the smallest RMSE of the tree-based methods. Boosting leads to a substantial improvement when compared to the simple decision tree; the boosted regressor has a RMSE of only 358 meV and $R^2 = 0.934$.

For the best model (nonlinear support vector machine), we plot the RMSE as a function of the training set size in Fig. 2. To generate this plot, we performed n-fold cross validation, where increasing n leads to an increase in the size of the training set. When the training set size is 124 (threefold cross validation, the RMSE is 336 meV; increasing the training set size by over 30% to 167 (10-fold cross validation), the RMSE is 283 meV.

The error of only 283 meV for the nonlinear support-vector machine is sufficiently small to make the model predictive in nature. This error is around the uncertainty in the band gaps for high-level first-principles calculations. Hybrid functionals rely on choosing a mixing parameter that affects the calculated gap, whereas the results from non-self-consistent GW calculations are sensitive to the choice of starting wave functions [36]; the error in the calculated gaps based on these approaches is typically 0.1–0.3 eV. Therefore, our fitted model provides a

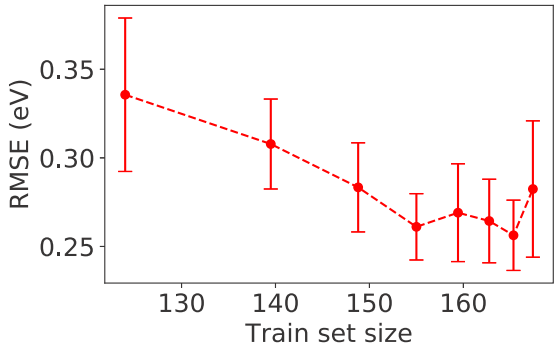

FIG. 2. For a nonlinear support vector machine, the root mean squared error (RMSE) is plotted for the band gap predictions as a function of training set size. The error bars represent the standard deviations in the RMSE from n-fold cross validation.

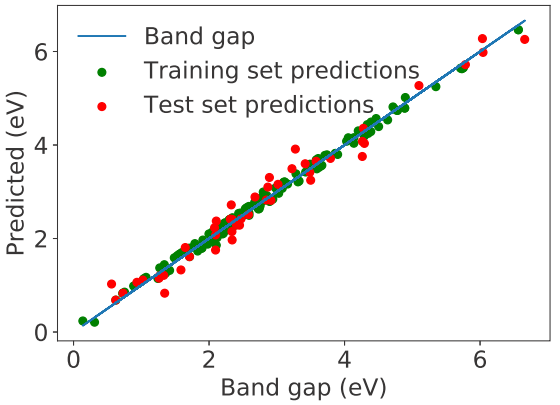

FIG. 3. Machine-learning predictions based on a support-vector regressor with a radial bias kernel. Predictions for the training set (green) circles and test set (red circles) are compared with the HSE calculated values (blue line).

degree of accuracy as good as the input band gaps calculated from first principles.

To visualize the accuracy of our band gap predictions, we plot each predicted gap as a function of the calculated HSE gap in Fig. 3, using the optimized support-vector machine with a radial bias kernel. We partitioned the HSE-calculated band gaps and features for the 184 compounds into a training and test set; approximately 75% of the data points are used to train the optimized machine learning model, and 25% are kept for testing. Figure 3 shows that the model provides highly accurate predictions for both the training and test set, when compared to the HSE calculated values.

B. Direct-indirect classifier

Based on our HSE calculations, of the 184 gaps used for fitting, 78 were found to be direct band gaps, and 108 were indirect. The classifier is trained on the simple 12-dimensional feature space that was used in regression. The model achieves an accuracy score of 73%.

To provide better predictions, we preform feature engineering. We attempted to construct differences, means, and standard deviations from the features, as was implemented previously [17]; however, this did not improve classifier performance. Improved predictions were achieved by constructing polynomial combinations of the original features in the 12-dimensional feature space. For second-order polynomial combinations, the accuracy of the classifier is increased to 83%. Using third-order polynomial features leads to a reduction in the accuracy to 81%; increasing the degree of the polynomial further led to a more dramatic reduction in classifier accuracy, suggesting overfitting.

To address the issue of overfitting, while still having the advantage of keeping some higher-order terms, we use the feature-selection method with recursive feature extraction. In this way, the high-dimensional polynomial feature space is pruned to a small subset of features that have the highest weighting in determining the outcome. Using feature extraction by fitting the classifier with third-order polynomial features, the accuracy score is increased to 89%; the optimum number of features is 30. Our optimized binary classifier is

TABLE II. Accuracy score for the direct-indirect classifier using logistic regression. Results are shown for different models with varied complexity in the feature space: (i) the simple 15-dimensional (15D) feature space described in Sec. II C 3, (ii) second order polynomial combinations (2nd PF) of the 15D set, (iii) third order polynomial features (3rd PF), and (iv) third order polynomial features plus feature extraction (3rd PF + FS).

Feature space	Accuracy score
15D	73%
2nd PF	83%
3rd PF	81%
3rd PF + FS	89%

described by the following metrics for classification performance: $precision = 0.88$; $recall = 0.91$; $f1 = 0.89$. (See Table II.)

C. Predicted results

1. Band gaps

With our fitted models, the band gaps of all 1568 materials are predicted. In Fig. 4, the band gap distributions are presented for the oxides, sulfides, selenides, and tellurides. Oxides typically have larger band gaps with a mean band gap $E_g^{av} = 3.82$ eV; however, the distribution of the band gaps is over a very wide energy range, with a standard deviation of $\sigma = 1.47$ eV. The trend moving down the Periodic Table for the anions is for smaller band gaps and a more localized distribution. For the tellurides, $E_g^{av} = 1.78$ eV and $\sigma = 0.68$ eV.

The direct-indirect predictions are shown in Fig. 5. Over all materials studied, 70% are found to have indirect band gaps and 30% are direct-gap materials. The percentage of materials that were direct or indirect was anion dependent; however, there was no clear systematic trend.

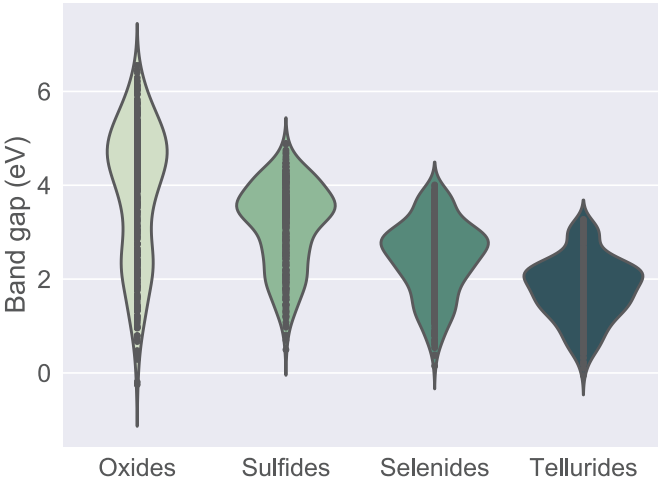

FIG. 4. Violin plot for the predicted band gap distributions for the oxides, sulfides, selenides, and tellurides. The width of each distribution at a given energy indicates the number of materials with a band gap around that energy.

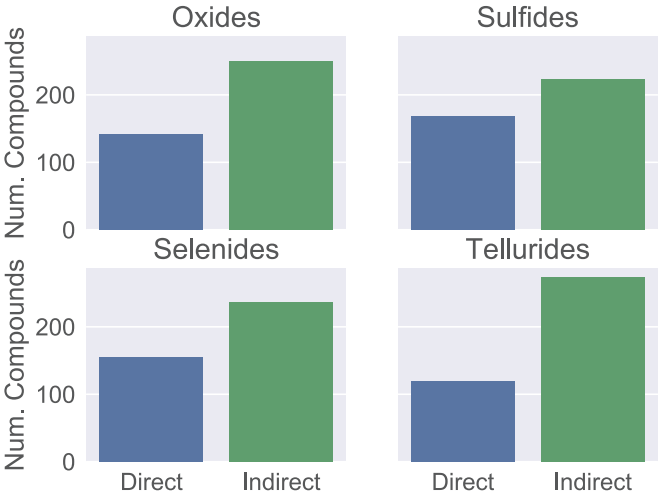

FIG. 5. Direct-indirect distributions for the band gaps of the oxides, sulfides, selenides, and tellurides.

Materials with a band gap in the range 0.5–2.5 eV are suitable solar absorbers [6]; of the 1568 materials studied, 717 had a band gap in this range. For optimum photovoltaic device performance, band gaps around 1.5 eV are optimum [6]. We have identified 242 materials with a band gap in the "*optimum range*" of 1.2–1.8 eV. The band gap properties of all 1568 kesterite I_2-II-IV-VI_4 compounds are tabulated in the Supplemental Material [37].

2. Material stability

In order to guide further experimental and theoretical work, we have also assessed the stability of the 242 compounds predicted to have band gaps in the optimum range of 1.2–1.8 eV. The stability was assessed by computing the enthalpy of formation for each compound, and calculating the energy of decomposition into other phases. This was achieved by making use of the Materials Project database [13], which contains the enthalpies of formation for hundreds of thousands of materials. In this way, we can determine whether a material is stable, metastable, or not stable.

Of the 242 compounds, 25 were found to be the ground state for that stoichiometry with respect to the Materials Project database; i.e., these materials are expected to be stable. An additional nine materials had an energy above hull of <0.1 eV/atom, and are expected to be metastable [38]. We therefore predict that 34 of these kesterites with a band gap in the optimum range should be synthesizable.

3. Band gap verification

As a final step, we verify the machine-learned band gaps of these 34 stable compounds using first-principles calculations with the HSE functional. In the end, we find that 25 of these materials actually had a band gap with optimum range of 1.2–1.8 eV. In Table III, the band gap properties for these 25 materials are presented. We indicate the magnitude of the fundamental gap, as well as the direct-indirect character. Materials that are the ground state for that stoichiometry are

TABLE III. Predicted properties for materials with band gaps in the "*optimum range*" of 1.2–1.8 eV. The magnitude of the fundamental gap (E_g), and the direct-indirect (Dir./Indir.) character of the gap are presented. The stability is also indicated; for materials that are metastable, the energy above hull (per atom) is indicated.

Material	E_g (eV)	Dir./Indir.	Energy above hull (eV)	Stability
$Li_2BeGeTe_4$	1.419	Direct	0	Stable
$Li_2BeSnTe_4$	1.611	Direct	0	Stable
$Rb_2BeSnTe_4$	1.692	Direct	0	Stable
$Rb_2HgTiSe_4$	1.751	Indirect	0	Stable
$Cs_2HgTiSe_4$	1.753	Indirect	0	Stable
$Cu_2BeSiTe_4$	1.251	Indirect	0	Stable
$Cu_2BeGeSe_4$	1.210	Indirect	0	Stable
$Cu_2MgSiTe_4$	1.272	Indirect	0	Stable
$Cu_2SrSiSe_4$	1.793	Indirect	0	Stable
$Cu_2ZnSiSe_4$	1.751	Direct	0	Stable
Cu_2ZnSnS_4	1.238	Direct	0	Stable
$Cu_2CdSiSe_4$	1.534	Direct	0	Stable
$Ag_2BeSiTe_4$	1.527	Indirect	0	Stable
$Ag_2BeGeSe_4$	1.489	Direct	0	Stable
$Ag_2MgSiTe_4$	1.591	Direct	0	Stable
$Ag_2MgGeSe_4$	1.322	Direct	0	Stable
$Ag_2SrSiTe_4$	1.543	Indirect	0	Stable
$Ag_2ZnSiSe_4$	1.787	Direct	0	Stable
$Ag_2CdSiSe_4$	1.640	Indirect	0	Stable
$Ag_2HgSiSe_4$	1.218	Direct	0	Stable
$Cs_2BeSnTe_4$	1.708	Direct	0.004	Metastable
$Na_2BeSnTe_4$	1.783	Direct	0.014	Metastable
$Ag_2SrSnSe_4$	1.272	Direct	0.018	Metastable
$Ag_2CaSiTe_4$	1.720	Direct	0.061	Metastable
Cu_2BeSnS_4	1.657	Direct	0.086	Metastable

indicated to have an energy above hull of 0 eV. For materials that are metastable, the energy above hull is indicated.

The 25 materials presented in Table III are largely unexplored. We encourage other researchers, both theoretical and experimental, to use the results in this manuscript as a guide in the materials engineering of kesterite I_2-II-IV-VI_4 semiconductors.

IV. CONCLUSIONS

We have determined the band gap properties of 1568 kesterite I_2-II-IV-V_4 semiconductors using a combination of first-principles calculations and machine learning. By performing explicit hybrid-functional calculations on a subset of 200 compounds, we trained machine learning models to predict the magnitude and character of the fundamental gap. A trained machine learning regressor based on a support-vector machine could predict the magnitude of the gap with a RMSE of 283 meV; a direct-indirect classifier was fit using logistic regression, and has an accuracy of 89%. Our predictions identify 242 materials with a band gap in the *optimum* range of 1.2-1.8 eV, and we expect that 34 of these materials are synthesizable; 25 of these materials actually had a band gap in the range of 1.2–1.8 eV, as verified using first-principles calculations with the HSE functional. These results will be useful in the materials engineering of solar absorbers for photovoltaic devices.

ACKNOWLEDGMENT

We gratefully acknowledge computational resources provided by the Australian National Computational Infrastructure (NCI) and support from the Australian Research Council.

[1] Z.-H. Cai, P. Narang, H. A. Atwater, S. Chen, C.-G. Duan, Z.-Q. Zhu, and J.-H. Chu, Chem. Mater. **27**, 7757 (2015).
[2] S. Chen, X. G. Gong, A. Walsh, and S.-H. Wei, Phys. Rev. B **79**, 165211 (2009).
[3] K. Ito and T. Nakazawa, Jpn. J. Appl. Phys. **27**, 2094 (1988).
[4] H. Katagiri, K. Jimbo, S. Yamada, T. Kamimura, W. S. Maw, T. Fukano, T. Ito, and T. Motohiro, Appl. Phys. Express **1**, 041201 (2008).
[5] S. Siebentritt and S. Schorr, Prog. Photovoltaics **20**, 512 (2012).
[6] S. Rühle, Solar Energy **130**, 139 (2016).
[7] J. Nelson, *The Physics of Solar Cells* (World Scientific, Singapore, 2003).
[8] S. Chen, J.-H. Yang, X. G. Gong, A. Walsh, and S.-H. Wei, Phys. Rev. B **81**, 245204 (2010).
[9] P. Mori-Sánchez, A. J. Cohen, and W. Yang, Phys. Rev. Lett. **100**, 146401 (2008).
[10] J. Heyd, J. E. Peralta, G. E. Scuseria, and R. L. Martin, J. Chem. Phys. **123**, 174101 (2005).
[11] M. Shishkin and G. Kresse, Phys. Rev. B **75**, 235102 (2007).
[12] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli *et al.*, Comput. Mater. Sci. **58**, 227 (2012).
[13] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder *et al.*, APL Mater. **1**, 011002 (2013).
[14] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM **65**, 1501 (2013).
[15] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, Sci. Rep. **3**, 2810 (2013).
[16] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, J. Chem. Theory Comput. **11**, 2087 (2015).
[17] J. Lee, A. Seko, K. Shitara, K. Nakayama, and I. Tanaka, Phys. Rev. B **93**, 115104 (2016).
[18] G. Pilania, A. Mannodi-Kanakkithodi, B. Uberuaga, R. Ramprasad, J. Gubernatis, and T. Lookman, Sci. Rep. **6**, 19375 (2016).
[19] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, npj Comput. Mater. **2**, 16028 (2016).
[20] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).
[21] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[22] P. E. Blöchl, Phys. Rev. B **50**, 17953 (1994).
[23] 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).
[24] M. J. Lucero, T. M. Henderson, and G. E. Scuseria, J. Phys.: Condens. Matter **24**, 145504 (2012).
[25] J. Heyd, G. Scuseria, and M. Ernzerhof, J. Chem. Phys. **118**, 8207 (2003).
[26] J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. **124**, 219906 (2006).

[27] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[28] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).
[29] W. Setyawan and S. Curtarolo, Comput. Mater. Sci. **49**, 299 (2010).
[30] S. Chen, W.-J. Yin, J.-H. Yang, X. Gong, A. Walsh, and S.-H. Wei, Appl. Phys. Lett. **95**, 052102 (2009).
[31] A. J. Smola and B. Schölkopf, Stat. Comput. **14**, 199 (2004).
[32] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees* (CRC Press, Boca Raton, FL, 1984).
[33] A. Liaw and M. Wiener, R News **2/3**, 18 (2002).
[34] J. Elith, J. R. Leathwick, and T. Hastie, J. Anim. Ecol. **77**, 802 (2008).
[35] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, Phys. Rev. Lett. **117**, 135502 (2016).
[36] M. van Schilfgaarde, T. Kotani, and S. Faleev, Phys. Rev. Lett. **96**, 226402 (2006).
[37] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevMaterials.2.085407 for predicted band gap properties for all 1568 kesterite compounds considered in this study.
[38] W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. Persson, and G. Ceder, Sci. Adv. **2**, e1600225 (2016).