Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques

Joohwi Lee,[1,*] Atsuto Seko,[1,2,3] Kazuki Shitara,[1,4] Keita Nakayama,[1] and Isao Tanaka[1,2,3,4]
[1]*Department of Materials Science and Engineering, Kyoto University, Kyoto 606-8501, Japan*
[2]*Elements Strategy Initiative for Structure Materials (ESISM), Kyoto University, Kyoto 606-8501, Japan*
[3]*Center for Materials Research by Information Integration, National Institute for Materials Science (NIMS), Tsukuba 305-0047, Japan*
[4]*Nanostructures Research Laboratory, Japan Fine Ceramics Center, Nagoya 456-8587, Japan*

(Received 4 September 2015; revised manuscript received 1 December 2015; published 1 March 2016)

Machine learning techniques are applied to make prediction models of the G_0W_0 band gaps for 270 inorganic compounds using Kohn-Sham (KS) band gaps, cohesive energy, crystalline volume per atom, and other fundamental information of constituent elements as predictors. Ordinary least squares regression (OLSR), least absolute shrinkage and selection operator, and nonlinear support vector regression (SVR) methods are applied with two levels of predictor sets. When the KS band gap by generalized gradient approximation of Perdew-Burke-Ernzerhof (PBE) or modified Becke-Johnson (mBJ) is used as a single predictor, the OLSR model predicts the G_0W_0 band gap of randomly selected test data with the root-mean-square error (RMSE) of 0.59 eV. When KS band gap by PBE and mBJ methods are used together with a set of predictors representing constituent elements and compounds, the RMSE decreases significantly. The best model by SVR yields the RMSE of 0.24 eV. Band gaps estimated in this way should be useful as predictors for virtual screening of a large set of materials.

DOI: 10.1103/PhysRevB.93.115104

I. INTRODUCTION

The band gap is a simple but important parameter to characterize semiconductors and insulators for optical and electronic applications [1]. To explore a material with a desired property from a large set of compounds, it can be a good starting point to know the band gap of all compounds in a library. However, experimental data for band gaps are still limited. Accurate measurement of the band gap requires high quality single crystals. However, they are often hard to synthesize. By virtue of recent advances in computational power and techniques, the first-principles calculation becomes a common approach to estimate the band gap of a large number of compounds. Actually, the band gap has been already computed for over 40,000 compounds, and they are included in several open-access databases [2–4]. Most of them were computed by the density functional theory (DFT) calculation [5], with the generalized gradient approximation (GGA) of Perdew-Burke-Ernzerhof (PBE) [6] exchange-correlation functional. The Kohn-Sham gap (KS gap), namely, the difference between lowest unoccupied and highest occupied eigenvalues is typically used as an approximated band gap. However, there is a well-known drawback; the KS gap by GGA underestimates the band gap significantly. As an example, Fig. 1 shows the comparison of the KS gap by GGA-PBE and experimental band gap for 32 crystals included in our data set that will be shown later. The root-mean-square (RMS) difference between PBE and experimental band gaps is as large as 2.25 eV.

A practical remedy for the underestimation is the use of a new exchange-correlation functional, such as the modified Becke-Johnson (mBJ) functional by Tran and Blaha [7]. Although the mBJ functional has been successful for evaluating the band gap in a number of compounds, there seems to be small but systematic inconsistency between the mBJ KS gap and the experimental band gap. As shown in Fig. 1, the deviation between the mBJ KS gap and the experimental band gap is 0.73 eV for the 32 crystals, as an average. Despite the improvement from PBE, researchers often need to evaluate the band gap more accurately. There are at least two different approaches for the improvement. One is so called the delta self-consistent-field (ΔSCF) method to evaluate the band gap from the differences in total energies [8]. The other uses hybrid functional methods [9] or GW calculations based on the many-body perturbation theory [10]. The downside of the latter methods is their high computational costs. It is still difficult to establish a large database of band gap by such methods. Therefore, an alternative way to accurately estimate the band gap using computationally affordable methods is desired.

Machine learning methods have been used for a number of purposes in condensed matter physics and materials science. The data-driven approaches to estimate a physical quantity are useful, especially when a target property cannot be directly computed by the GGA level calculations. A prediction model can be constructed by a regression method, taking the GGA level data set and other fundamental information of constituent elements and crystal structure as predictors. Such methods have been applied to estimate a wide range of material properties, such as the melting temperature [11,12], the ionic conductivity [13,14], the potential energy surface [15–17], the phase stability [18], the formation energy of crystals [19–21], the atomization energy of molecules [22–24], and the density of states [25]. On the prediction of the band gap, a few applications of regression methods have been reported [26–30]. Setyawan *et al.* estimated a relationship between the PBE KS gap and the experimental band gap by ordinary least squares regression (OLSR) from a data set composed of about 100 compounds established on the database AFLOWLIB [2], including both direct and indirect band gaps [26]. Dey *et al.* predicted the direct band gap of about 200 ternary chalcopyrite compounds from 28 experimental

[*] Corresponding author: lee.joohwi@gmail.com

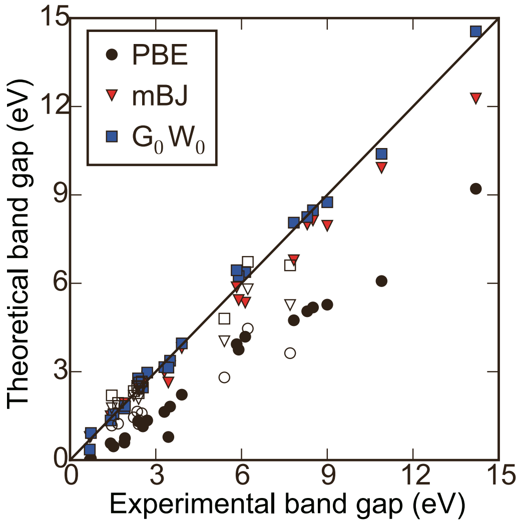

FIG. 1. Comparison of theoretical and experimental band gaps. The closed and open symbols indicate the direct (22 compounds) and indirect (10 compounds) band gaps.

band-gap observations by OLSR, sparse partial least square regression, and least absolute shrinkage and selection operator (LASSO) methods with predictors, such as valence, atomic number, melting point, electronegativity, and pseudopotential radii of each element [27]. Gu *et al.* applied support vector regression (SVR) and artificial neural networks (ANNs) to predict experimental band gaps of 25 binary and 31 ternary compounds with some element-specific predictors [28]. Montavon *et al.* [29] and Ramakrishnan *et al.* [30] used ANNs and a Δ-machine learning approach [24] with a Coulomb matrix descriptor to predict the levels of highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO) of GW calculations and the differences between DFT band gaps and time-dependent DFT excitation energies of more than a few thousand organic molecules, respectively. Despite these pioneering works, there is still a plenty of room for improvement by selecting better regression techniques and predictors for the machine learning. In this paper, we construct prediction models of the band gap using different kinds of predictors and regression techniques. First, we estimate relationships between the band gaps obtained by different levels of approximations in the first-principles calculation. We set a target property to the quasiparticle gap (QP gap) for 270 binary and ternary compounds that is obtained by the G_0W_0 calculation with Heyd-Scuseria-Ernzerhof hybrid functional (HSE06) instead of the experimental band gap. As shown in Fig. 1, the G_0W_0 band gap is, averagely, the closest to the experimental band gap. Second, prediction models for the QP gap by the G_0W_0 calculation, E_g (G_0W_0), are constructed from DFT KS gaps, cohesive energy, crystalline volume per atom, and fundamental information of constituent elements using OLSR, LASSO, and nonlinear SVR.

II. METHODOLOGY

A. Regression methods

The relationship between a target property and predictors can be obtained by regression methods. The reliability of the estimation depends on the quality of the training set, selection of predictors, and regression method. For a given training set, it is essential to select "good" predictors and regression methods. In the present paper, we compare the OLSR, LASSO, and nonlinear SVR regression methods. In the OLSR regression, regression coefficients of predictors, $\boldsymbol{\beta}$, are determined from n observations by minimizing the following minimization function $L(\boldsymbol{\beta})$, given by

$$L(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \qquad (1)$$

where $\| \|_2$ denotes the L2 norm. $\boldsymbol{X}$ and $\boldsymbol{y}$ are $(n \times p)$ predictor matrix and n-dimensional vector of target property for training data, respectively. The LASSO regression enables us not only to provide a solution for linear regression but also to obtain a sparse representation with a small number of nonzero regression coefficients [31]. The LASSO minimization function is defined as

$$L(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \qquad (2)$$

where $\| \|_1$ denotes the L1 norm. The parameter λ controls the trade-off relationship between sparsity and accuracy. The LASSO regression is useful when the number of candidate predictors exceeds or is as large as the number of training data.

Meanwhile, nonlinear regressions are flexible to express complex functions. The SVR [32] is a nonlinear regression method based on the kernel trick [33]. The SVR is a version of a support vector machine for regression. The kernel function maps the original feature space of the considered data into a high-dimensional space, where the learning task is simplified. Here, we use the SVR with the Gaussian kernel function. The Gaussian kernel function between two data points $\boldsymbol{x}$ and $\boldsymbol{x}'$ with a parameter σ^2 is defined as

$$k(\boldsymbol{x},\boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right). \qquad (3)$$

The SVR is performed using the e1071 package [34] implemented in R software [35]. Internal parameters in the SVR are optimized by minimizing the tenfold cross validation (CV) score for the training data set. The prediction power of regression models is evaluated by the RMS error (RMSE) of test data set.

B. Computational detail of first-principles calculations

We made a first-principles data set composed of 270 inorganic compounds. They include 156 AX binary compounds of I-VII, II-VI, III-V, and IV-IV elements with four kinds of the crystal structures, i.e., wurtzite- (WZ), cesium chloride- (CC), zincblende- (ZB), and rocksalt-type (RS). Only those compounds exhibiting positive E_g(PBE) and E_g(mBJ) are adopted in the data set. The data set also includes 30 nonequiatomic binary and 84 ternary ionic compounds that show positive E_g(PBE). They were selected randomly from materials projects database [3] with the following conditions: (1) The compound is composed only of nontransition metal elements. (2) The unit cell is composed of less than 25 atoms. (3) E_g(PBE) is positive. All of these compounds were subjected to GGA (PBE), mBJ, and the G_0W_0 calculations. In addition, the cohesive energy and volume per atom were calculated by the PBE functional to be used as a predictor for building prediction models of E_g (G_0W_0).

All of first-principles calculations were performed by the projector-augmented wave (PAW) method [36,37] implemented in the Vienna *Ab initio* Simulation Package (VASP) [38,39]. The cutoff energy was set to 500 eV for PBE and mBJ calculations. Before calculating band gaps, the structure optimization was performed using the PBE functional revised for solids (PBEsol) [40] until residual forces acting on atoms reached below 0.005 eV Å^{-1}. For k-space sampling, $k_1 \times k_2 \times k_3$ Γ-centered meshes were used, where k_n ($n = 1,2,3$) was prepared as the near natural number of 40 per lattice parameter (1 Å^{-1}) to each direction. For the G_0W_0 calculation with the HSE06 functional, a set of pseudopotentials updated for *GW* calculations was used. The cutoff energy was set to

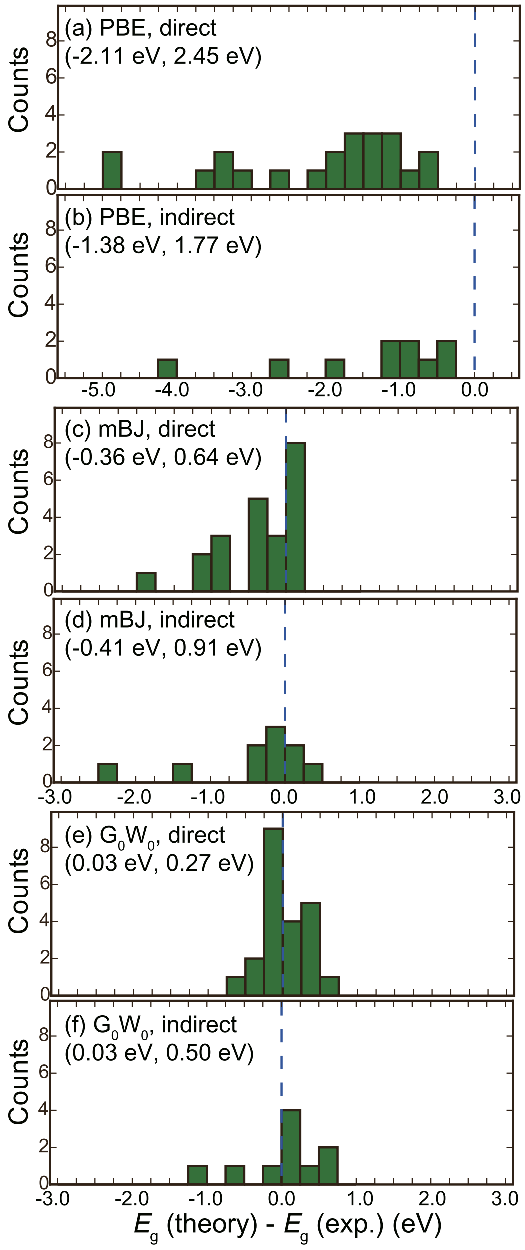

FIG. 2. Histogram of distribution of differences for theoretical and experimental band gaps. Theoretical band gaps are obtained for direct-gap compounds by (a) PBE, (c) mBJ, and (e) G_0W_0 and by (b) PBE, (d) mBJ, and (f) G_0W_0, respectively, for indirect-gap compounds. Dashed vertical line corresponds to E_g(theory) = E_g (exp). Values in parentheses are mean and RMS differences.

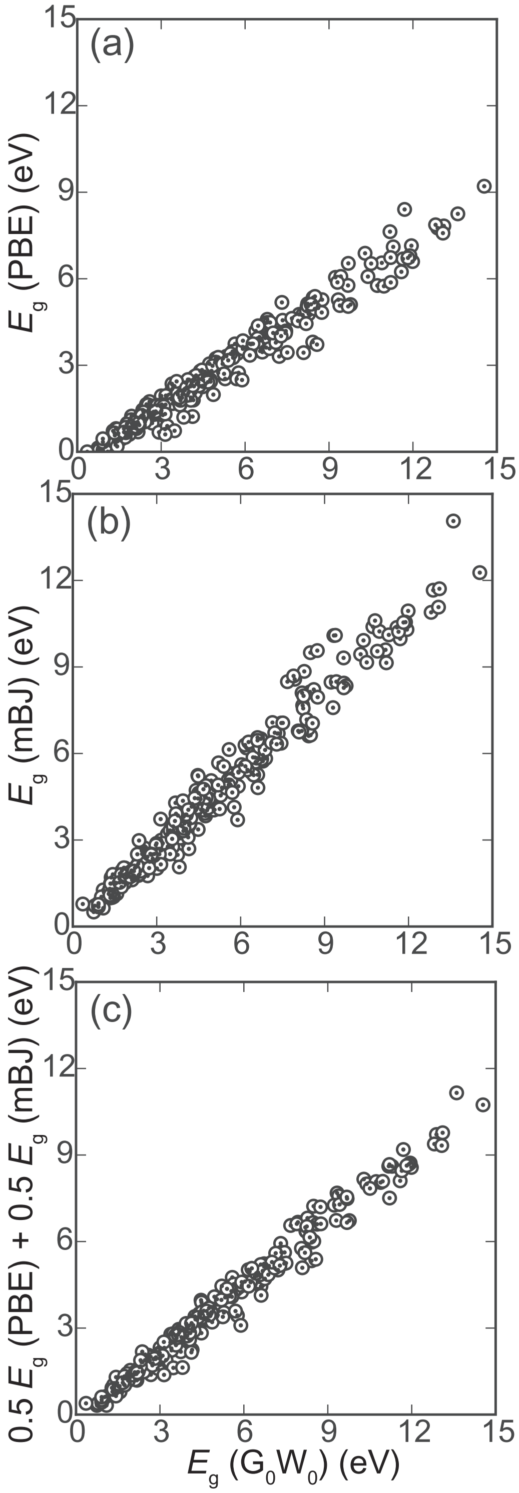

FIG. 3. Relationships between the QP gap by the G_0W_0 calculation with HSE06 functional and KS gap of (a) E_g(PBE), (b) E_g (mBJ), and (c) the average of E_g(PBE) and E_g (mBJ).

600 eV. The number of k-points was reduced to the half of the DFT calculations for computational efficiency.

III. RESULTS AND DISCUSSION

A. Comparison of theoretical band gaps

Among 156 AX binary compounds, we found experimental band gaps of 32 compounds, as shown in the Appendix. They are compared to the theoretical band gaps in Fig. 1. The tendency of the band-gap accuracy with respect to methods is similar to previous reports [7,10]. Figure 2 shows the histogram of differences between the theoretical and experimental band gaps. The mean value of the difference between the PBE KS gap and experimental gap is −2.11 and −1.38 eV for direct and indirect compounds, respectively. Their RMS differences are 2.45 and 1.77 eV, respectively. Using the mBJ functional, theoretical band gaps are notably improved. However, a tendency to underestimate the band gap can still be noted by the mBJ calculations. The G_0W_0 calculations show a much better description of the band gaps not only for the mean value but also for the RMS difference. Since the number of experimental data is limited, we will use the G_0W_0 QP gap as the target property in the present paper. Considering the fact that the experimental band gaps by different groups are sometimes scattered, especially for the indirect band gaps, the present results may be applicable to estimate the real band gaps.

B. Correction of band gaps of PBE and mBJ

Figure 3(a) shows the relationship between $E_g\,(G_0W_0)$ and $E_g\,$(PBE) for 270 compounds. Direct and indirect band-gap materials are shown together. A similar plot is made for $E_g\,(G_0W_0)$ versus $E_g\,$(mBJ) in Fig. 3(b). Both of them exhibit almost the linear dependence. Prediction models of $E_g\,(G_0W_0)$ are then made by the OLSR only from E_g(PBE) and $E_g\,$(mBJ). To estimate the prediction error of the models, the whole data set is randomly divided into a training data composed of three quarters of the whole data set; the rest is used as a test data. The random selection of the training data set is repeated for 100 times. The prediction error is estimated as the RMSE for the test data and tenfold CV score. We also evaluate the mean absolute percentage error (MAPE) for the test data defined as

$$\frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}}\left|\frac{y_i-\hat{y}_i}{y_i}\right|\times 100 \qquad (4)$$

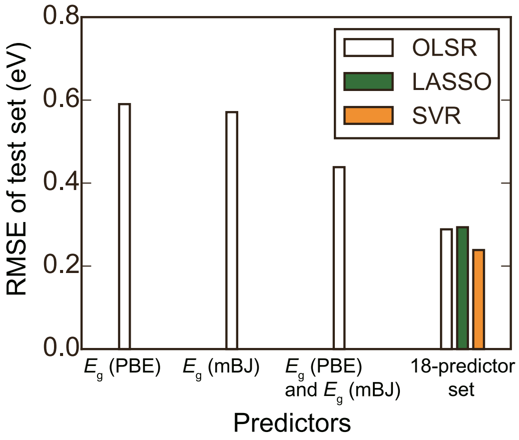

FIG. 4. Dependence of the prediction error on the predictor set. Black open bars and green and orange closed bars denote the RMSEs by the OLSR, LASSO, and SVR models, respectively. The values are given in Table I.

where y_i, $\hat{y}_i$, and n_{test} denote the observed target property for data i, predicted target property for data i, and the number of test data, respectively. The RMSE, CV score, and MAPE are averaged over 100 trials. Table I summarizes the RMSE, CV score, and MAPE of the OLSR models constructed by E_g(PBE) and $E_g\,$(mBJ). The OLSR model with a single predictor, E_g(PBE) or $E_g\,$(mBJ), shows RMSE of 0.59 and 0.57 eV, respectively.

When both E_g(PBE) and $E_g\,$(mBJ) are used as predictors, the OLSR model shows the RMSE of 0.44 eV, which is much smaller than those of the OLSR models with single predictors. The same behavior can be seen in the CV score and MAPE. A linear model is obtained as $E_g\,(G_0W_0) = 0.75\,E_g(\text{PBE}) + 0.56\,E_g\,(\text{mBJ}) + 0.36$ in the unit of eV. The physics behind the improvement of using two KS gaps is not clear at the moment. However, the correlation coefficient for the linear plot is increased when both of E_g(PBE) and $E_g\,$(mBJ) are used as predictors instead of a single KS gap, which is the phenomenological reason for the improvement. The relationship between $E_g\,(G_0W_0)$ and the average of E_g(PBE) and $E_g\,$(mBJ) is shown in Fig. 3(c).

C. Comparison of different prediction models

To further improve the prediction models for $E_g\,(G_0W_0)$, some additional predictors are examined. For the purpose of

TABLE I. Prediction errors for the OLSR, LASSO, and SVR models with E_g(PBE), $E_g\,$(mBJ), and the 18-predictor set. The standard deviation of errors between 100 trials is in parentheses.

Regression methods	Predictors	Average number of selected predictors	RMSE (eV)	CV score (eV)	MAPE (%)
OLSR	E_g(PBE)		0.59 (0.06)	0.59 (0.02)	10.66
OLSR	$E_g\,$(mBJ)		0.57 (0.07)	0.57 (0.02)	10.48
OLSR	E_g(PBE), $E_g\,$(mBJ)		0.44 (0.04)	0.44 (0.02)	8.12
OLSR	Selected from the 18-predictor set	13.3	0.29 (0.03)	0.29 (0.01)	6.50
LASSO	Selected from the 18-predictor set	14.4	0.29 (0.03)	0.29 (0.01)	6.43
SVR	18-predictor set		0.24 (0.03)	0.24 (0.01)	5.12

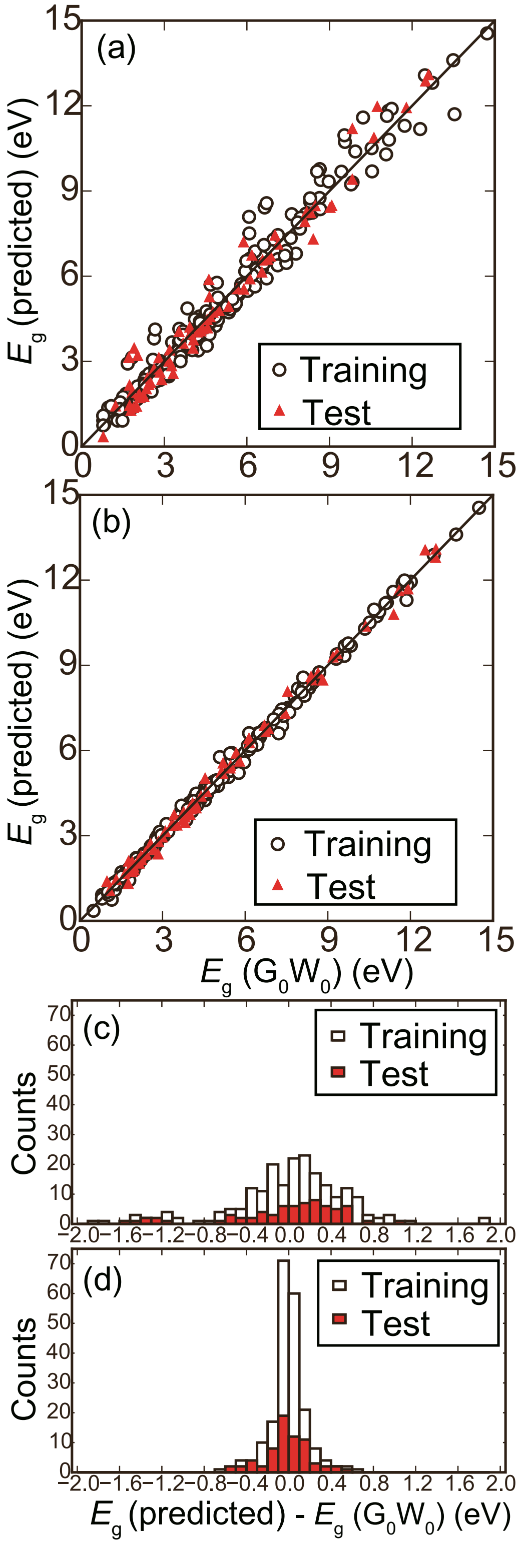

FIG. 5. Comparison of E_g (G_0W_0) method and E_g (predicted). (a) and (c) The OLSR model with E_g(PBE) as a single predictor. (b) and (d) The SVR model with the 18-predictor set. A randomly chosen result from 100 trials is shown.

screening a large library, the use of a simple predictor set included in a library, or that can be easily made by combining physical quantities in the library, is essential. Alternatively, the predictors may be routinely computed by affordable DFT calculations. In the present paper, we use quantities obtained by GGA-PBE calculations, i.e., the cohesive energy, E_{coh}, and crystalline volume per atom, V, in addition to E_g(PBE) and E_g (mBJ) as predictors. The other seven fundamental variables related to constituent elements are also taken as candidates of predictors. They are the absolute value of the formal ionic charge $|n|$, period p in the periodic table, atomic number Z, atomic mass m, van der Waals radius r^{vdW}, electronegativity χ, and the first ionization energy I [41]. For a given compound, the element-specific predictors are made into average forms, $\langle\xi\rangle$, and standard deviation forms, σ_ξ, namely,

$$\langle\xi\rangle = \sum_k^N x_k \xi_k \tag{5}$$

$$\sigma_\xi = \sqrt{\sum_k^N x_k(\xi_k - \langle\xi\rangle)^2} \tag{6}$$

where x_k and ξ_k are the composition and fundamental variables of each constituent element, respectively. All of them are included in the 18-predictor set.

Using the 18-predictor set, prediction models of E_g (G_0W_0) are constructed by three regression methods, i.e., OLSR, LASSO, and SVR. Note that the training data is standardized before performing the regressions. For the OLSR, the stepwise optimization based on the Akaike information criterion (AIC) [42] is used to select an appropriate set of predictors and avoid the overfitting. Figure 4 and Table I show the prediction errors of the OLSR, LASSO, and SVR models. As described before, when both E_g(PBE) and E_g (mBJ) are used as predictors, the OLSR model provides the RMSE of 0.44 eV. The OLSR model with stepwise regression using the AIC gives the RMSE of 0.29 eV, which is much improved compared to the OLSR models with E_g(PBE) and E_g (mBJ). The result by the LASSO model is almost the same as that of the present OLSR. This is quite natural since the average numbers of selected predictors are almost the same as 13.3 for the OLSR and 14.4 for the

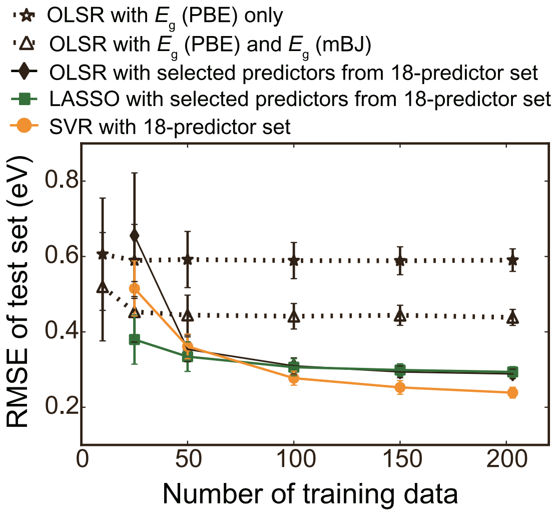

FIG. 6. Dependence of RMSE on the number of training data for five models. Error bars show standard deviations for 300 trials.

LASSO. The SVR model provides the RMSE of 0.24 eV, which is better than the OLSR and LASSO models. The CV score and MAPE show the same tendency as the RMSE.

Figures 5(a) and 5(b) show the comparison of $E_g\,(G_0W_0)$ and E_g (predicted) by two models. They are randomly chosen from 100 trials. Compared to the OLSR model only with E_g(PBE) [Fig. 5(a)], deviation of the data from the diagonal straight line, where $E_g(\text{predicted}) = E_g\,(G_0W_0)$, is much smaller for the SVR model. For better viewing of the improvement of the prediction power, the distribution of the prediction error, $E_g(\text{predicted}) - E_g\,(G_0W_0)$, is shown in Figs. 5(c) and 5(d). The distribution of the prediction error for the test sets is similar, reflecting the same values of CV score and RMSE shown in Table I. This implies the absence of any serious overfitting problems.

Finally, we examine the dependence of the RMSE of the test data with respect to the number of training data. The RMSE is examined over 300 trials to obtain the average and standard deviation. The training and test sets are randomly chosen with the ratio of three to one from 270 compounds in each trial. Figure 6 shows the comparison of five models, i.e., OLSR model only with E_g(PBE), OLSR model with E_g(PBE) and E_g (mBJ), and three models with the 18-predictor set. It is interesting that the prediction powers of the first two OLSR models are saturated when the number is 50. On the other hand, the dependence is quite different when the 18-predictor set is used. The RMSE by the OLSR and LASSO models seem to be saturated at the number of about 200. But the RMSE by the SVR model still slowly decreases at about 200. The SVR model is expected to be improved with a larger number of training data. However, the number of training data should be chosen by considering the trade-off relationship between the computational costs and prediction power. The present model may already be useful for establishing a large database of band gap with reasonable accuracy.

IV. CONCLUSION

In this paper, we have developed the prediction models of the G_0W_0 band gaps for 270 inorganic compounds using KS band gaps, cohesive energy, crystalline volume per atom, and other fundamental information of constituent elements as predictors. The OLSR method is used with three levels of predictor sets. The LASSO and SVR methods are used only with the 18-predictor set. When E_g(PBE) is used as a single predictor, the OLSR model predicts the G_0W_0 band gap of a randomly selected test data with the RMSE of 0.59 eV. When E_g (mBJ) is additionally used together with the 18-predictor set, the RMSE decreases significantly. The best model by the nonlinear SVR yields the RMSE of 0.24 eV. In the prediction models, the serious overfitting problem does not take place. Once this type of well-corrected set of band gaps is established, it should be useful as predictors for virtual screening of a large set of materials.

ACKNOWLEDGMENTS

This paper was supported by Grant-in-Aid for Scientific Research (A) and Grant-in-Aid for Scientific Research on Innovative Areas “Nano Informatics” (Grant No. 25106005) from the Japan Society for the Promotion of Science (JSPS) and “Materials research by Information Integration“ Initiative (MI^2I) from Japan Science and Technology Agency. J.L. acknowledges Grant-in-Aid for International Research Fellow of JSPS (Grant No. 2604376) and JSPS fellowships.

APPENDIX: EXPERIMENTAL AND THEORETICAL BAND GAPS

Table II shows experimental and theoretical band gaps of 270 compounds used in this paper.

TABLE II. The band gaps of the compounds that are used in this paper.

Compound	Space group	E_g (exp.; eV)	Type[a]	E_g (PBE; eV)	E_g (mBJ; eV)	$E_g(G_0W_0$; eV)
Al_2CdTe_4	82		D	1.61	2.36	2.86
Al_2O_3	167		D	5.88	7.59	9.30
Al_2ZnS_4	227		D	2.50	3.58	3.99
Al_2ZnSe_4	82		D	2.13	3.24	3.77
AlAs	186		I	1.74	2.38	2.59
AlAs	216	2.23[b]	I	1.45	2.18	2.34
AlF_3	167		D	7.58	11.07	13.07
AlN	186	6.13[b]	D	4.19	5.34	6.38
AlN	216		I	3.31	4.81	5.33
AlN	225		I	4.59	5.73	6.79
AlP	186		I	1.94	2.70	3.00
AlP	216	2.45[c]	I	1.59	2.42	2.67
AlSb	186		D	0.99	1.54	1.79
AlSb	216	1.68[b]	I	1.24	1.84	1.94
$Ba(MgBi)_2$	164		D	0.41	0.99	1.34
BAs	186		I	1.11	1.71	1.93
BAs	216	1.46[b]	I	1.17	1.76	2.20
Be_2C	225		I	1.16	1.81	2.39
$BeCN_2$	122		D	3.85	5.34	5.92
BeO	186		D	7.63	9.59	11.18

TABLE II. (*Continued.*)

Compound	Space group	E_g (exp.; eV)	Type[a]	E_g (PBE; eV)	E_g (mBJ; eV)	$E_g(G_0W_0$; eV)
BeO	216		I	6.88	9.44	10.29
BeO	221		I	2.85	4.05	5.07
BeO	225		I	8.40	9.98	11.70
BeS	186		I	3.64	4.86	5.92
BeS	216		I	3.10	4.34	4.92
BeS	225		I	0.99	1.77	2.35
BeSe	186		I	3.24	4.15	4.95
BeSe	216		I	2.61	3.59	4.19
BeSe	225		I	0.00	0.64	1.09
BeTe	186		I	2.18	2.77	3.40
BeTe	216		I	1.95	2.64	3.17
BN	186		I	5.18	6.69	7.32
BN	216	6.22[b]	I	4.46	5.80	6.73
BN	225		I	1.24	1.76	2.68
BP	186		I	0.99	1.94	2.10
BP	216	2.40[b]	I	1.22	2.07	2.27
BSb	186		I	0.81	1.25	1.38
BSb	216		I	0.72	1.13	1.28
$Ca(CdAs)_2$	164		D	0.19	1.03	1.02
$Ca(CdP)_2$	164		D	0.80	1.66	1.67
Ca_3BiN	221		D	0.37	1.01	1.41
Ca_3PCl_3	221		D	1.79	2.69	4.15
Ca_3PN	221		D	0.80	1.62	2.12
$CaAlF_5$	15		D	7.15	10.30	11.95
$CaCl_2$	58		D	5.39	6.63	8.50
$CaCl_2$	136		D	5.35	6.61	8.43
CaF_2	225		I	7.11	10.11	11.30
$CaGeN_2$	122		D	2.65	4.15	4.25
$CaMg_2N_2$	164		D	1.98	3.33	3.59
CaO	186		D	3.24	5.68	5.21
CaO	216		I	3.38	6.14	5.59
CaO	221		I	2.74	4.14	5.77
CaO	225	7.70[d]	I	3.63	5.26	6.61
CaS	186		I	2.70	3.84	4.74
CaS	216		I	3.74	5.31	5.64
CaS	221		D	1.27	2.35	3.06
CaS	225		I	2.29	3.37	4.48
CaSe	186		I	2.39	3.26	4.09
CaSe	216		I	3.28	4.72	5.04
CaSe	221		D	0.69	1.69	2.17
CaSe	225		I	1.97	2.81	3.94
$CaSiO_3$	221		I	3.46	4.81	6.62
CaTe	186		I	1.87	2.49	3.75
CaTe	216		I	3.05	3.93	4.73
CaTe	221		I	0.13	0.50	0.75
CaTe	225		I	1.43	2.01	3.01
$CaZnF_4$	88		D	5.10	8.35	9.77
$Cd(GaTe_2)_2$	82		D	1.04	1.78	2.16
$Cd(InTe_2)_2$	82		D	0.84	1.52	1.87
$CdGa_2Se_4$	82		D	1.34	2.41	2.63
$CdIn_2Se_4$	82		D	0.94	1.87	2.07
$CdIn_2Te_4$	82		D	0.84	1.52	1.87
CdO	186		D	0.00	1.27	1.09
CdS	186		D	1.22	2.72	2.57
CdS	216	2.55[b]	D	1.14	2.63	2.46
CdS	225		I	0.24	1.70	1.37
CdSe	186		D	0.65	2.00	1.81
CdSe	216	1.90[b]	D	0.59	1.93	1.75

TABLE II. (*Continued.*)

Compound	Space group	E_g (exp.; eV)	Type[a]	E_g (PBE; eV)	E_g (mBJ; eV)	$E_g(G_0W_0$; eV)
$CdSiAs_2$	122		D	0.36	1.23	1.30
$CdSnF_6$	148		I	3.80	6.86	8.42
CdTe	186		D	0.79	1.84	1.92
CdTe	216	1.92[b]	D	0.75	1.79	1.84
Ga_2Se_3	9		D	1.04	1.92	2.42
Ga_2TeO_6	136		D	0.73	2.51	3.48
GaAs	186		D	0.48	1.48	1.52
GaAs	216	1.52[b]	D	0.46	1.51	1.56
GaN	186	3.50[b]	D	1.82	3.19	3.37
GaN	216	3.30[b]	D	1.64	2.99	3.16
GaN	225		I	0.60	2.03	1.86
GaP	186		D	1.36	2.27	2.38
GaP	216	2.35[b]	I	1.65	2.52	2.48
GaSb	186		D	0.06	0.72	0.79
GaSb	216	0.73[b]	D	0.05	0.81	0.92
Ge_3N_4	176		D	1.88	3.31	4.06
GeC	186		I	2.36	3.21	3.39
GeC	216		I	1.63	2.44	2.54
GeO_2	136		D	1.22	3.09	4.12
InN	186	0.69[b]	D	0.00	0.78	0.36
InP	186		D	0.64	1.55	1.47
InP	216	1.42[b]	D	0.57	1.49	1.34
K_2GeF_6	186		D	5.87	9.14	11.21
K_2MgO_2	60		D	2.50	4.89	4.76
K_2O	225		I	1.62	4.29	3.67
K_2S	225		I	2.28	4.72	4.41
K_2ZnTe_2	72		D	2.04	3.47	3.59
$KAlO_2$	92		D	3.75	6.22	6.67
KBO_2	167		D	4.12	7.07	7.14
$KCdF_3$	62		D	3.31	6.72	7.22
KCl	186		D	4.77	8.57	7.93
KCl	216		I	4.78	8.85	8.27
KCl	221		I	5.05	7.67	8.22
KCl	225	8.50[d]	D	5.18	8.13	8.48
KF	186		D	5.27	10.09	9.33
KF	216		I	5.07	10.10	9.39
KF	221		I	6.55	9.55	10.89
KF	225	10.90[d]	D	6.08	9.92	10.39
KI	186		D	3.73	6.19	6.21
KI	216		I	3.95	6.51	6.72
KI	221		I	3.40	5.27	5.72
KI	225		D	3.98	5.86	6.51
KLi_2As	59		D	0.68	1.84	1.68
KNaS	62		D	2.49	4.63	4.50
KNaSe	62		D	1.98	3.85	3.76
KZnP	194		D	0.81	2.04	1.82
Li_2CN_2	139		D	3.74	6.40	6.28
Li_2S	225		I	3.40	4.80	5.48
Li_2Se	225		I	2.98	4.11	4.84
Li_2Te	225		I	2.50	3.32	3.95
Li_3Sb	194		I	0.96	1.56	1.77
Li_4GeO_4	63		D	4.12	6.36	7.45
$LiAlTe_2$	122		D	2.44	3.04	3.55
$LiBF_4$	152		I	8.25	14.06	13.61
LiCaN	62		D	1.38	2.98	2.37
LiCaSb	62		D	0.62	1.03	1.46
LiCl	186		D	6.05	8.47	9.23
LiCl	216		D	6.08	8.49	9.42

TABLE II. (*Continued.*)

Compound	Space group	E_g (exp.; eV)	Type[a]	E_g (PBE; eV)	E_g (mBJ; eV)	$E_g(G_0W_0$; eV)
LiCl	221		I	4.56	6.70	7.34
LiCl	225		D	6.53	8.44	9.69
LiF	186		D	7.76	11.66	12.88
LiF	216		D	7.83	11.71	13.11
LiF	221		I	7.87	10.89	12.81
LiF	225	14.20[c]	D	9.21	12.27	14.55
$LiGaTe_2$	122		D	1.60	2.29	2.98
LiI	186		D	4.49	5.95	6.80
LiI	216		D	4.48	5.83	6.89
LiI	221		D	2.27	3.39	4.05
LiI	225		I	4.37	5.25	6.47
$LiInSe_2$	122		D	1.58	2.51	3.09
$LiInTe_2$	122		D	1.39	2.03	2.73
LiMgN	62		D	2.22	3.48	3.90
LiMgN	216		D	2.27	3.55	4.17
LiZnAs	216		D	0.50	1.59	1.71
Mg_3NF_3	221		D	3.58	6.12	6.87
$MgAl_2O_4$	227		D	5.11	7.17	8.36
MgF_2	136		D	6.80	10.55	11.89
MgF_2	205		D	6.69	10.22	11.64
$MgGeO_3$	15		I	2.53	4.65	5.70
MgO	186		D	3.49	5.70	6.54
MgO	216		D	3.61	5.81	6.75
MgO	221		I	2.53	4.55	5.29
MgO	225	7.83[c]	D	4.74	6.78	8.06
MgS	186		D	3.48	5.08	5.55
MgS	216		D	3.49	5.08	5.65
MgS	225	5.40[b]	I	2.80	4.03	4.80
MgSe	186		D	2.66	4.05	4.58
MgSe	216		D	2.66	4.04	4.66
MgSe	225		I	1.80	2.79	3.41
MgTe	186		D	2.50	3.56	4.19
MgTe	216		D	2.46	3.52	4.16
MgTe	225		I	0.44	1.13	1.57
Na_2O	225		D	1.98	4.43	4.87
Na_2S	225		D	2.43	4.47	4.77
Na_2Se	225		D	2.04	3.77	4.22
Na_2Te	225		D	2.05	3.41	3.97
Na_2ZnS_2	72		D	2.36	4.25	4.27
Na_3Sb	194		D	0.39	1.31	1.32
$NaAlO_2$	33		I	3.79	6.38	7.10
$NaCdF_3$	161		D	3.44	7.06	7.51
NaCl	186		D	4.88	8.10	8.20
NaCl	216		D	4.98	8.22	8.61
NaCl	221		I	4.19	7.06	7.45
NaCl	225	9.00[d]	D	5.27	7.95	8.75
NaF	186		D	5.77	10.39	10.73
NaF	216		D	5.74	10.23	10.96
NaF	221		I	6.24	10.37	11.59
NaF	225		D	6.60	10.94	11.99
NaI	186		D	3.61	5.61	6.03
NaI	216		D	3.86	5.84	6.58
NaI	221		I	2.35	4.25	4.58
NaI	225	5.90[d]	D	3.75	5.43	6.24
$NaInO_2$	166		I	1.79	3.88	4.06
$NaLi_2Sb$	225		I	0.66	1.37	1.74
NaLiS	129		D	3.05	4.91	5.20
NaMgAs	129		D	0.93	2.03	2.22

TABLE II. (*Continued.*)

Compound	Space group	E_g (exp.; eV)	Type[a]	E_g (PBE; eV)	E_g (mBJ; eV)	$E_g(G_0W_0$; eV)
$NaMgF_3$	62		D	6.72	10.54	11.82
NaSrAs	189		D	0.79	1.81	1.43
NaSrP	189		D	1.23	2.11	2.30
NaZnAs	129		D	0.32	1.49	1.36
$NaZnF_3$	62		D	3.72	7.05	8.57
NaZnP	129		D	0.87	1.92	2.09
Rb_2MgO_2	60		D	2.25	4.44	4.48
Rb_2O	225		I	1.31	3.73	3.14
Rb_2PbCl_6	225		D	1.20	2.06	3.81
Rb_2S	225		I	1.96	4.36	3.94
Rb_2Se	225		I	1.79	3.91	3.69
Rb_2SnCl_6	225		D	2.49	3.70	5.89
Rb_2Te	225		I	1.88	3.66	3.65
Rb_3GaO_3	12		D	2.42	4.76	4.67
$RbAlO_2$	227		I	3.35	5.56	6.17
RbCaAs	129		I	1.26	2.51	2.31
RbCl	186		D	4.63	8.49	7.67
RbCl	216		I	4.54	8.69	7.89
RbCl	221		I	5.13	7.56	8.24
RbCl	225	8.30[d]	D	5.05	8.00	8.25
RbF	186		D	4.96	9.50	8.49
RbF	216		I	4.83	9.57	8.74
RbF	221		D	6.53	9.16	10.51
RbF	225		I	5.77	9.31	9.68
RbI	186		D	3.75	6.29	6.16
RbI	216		I	3.85	6.55	6.60
RbI	221		I	3.57	5.37	5.92
RbI	225	5.83[d]	D	3.93	5.87	6.44
RbLiO	62		D	2.29	4.45	4.36
RbLiSe	129		I	2.29	4.05	4.13
SiC	186		I	2.28	3.33	3.47
SiC	216	2.42[b]	I	1.35	2.35	2.63
Sn_3N_4	227		D	0.19	1.10	1.43
SnO_2	136		D	0.60	2.15	3.15
Sr_2Si	62		D	0.34	0.68	0.92
Sr_2SnO_4	64		I	2.70	4.08	5.25
$SrBe_3O_4$	190		D	4.12	6.45	7.00
SrCaGe	62		D	0.45	0.76	0.92
$SrCN_2$	166		I	3.17	5.55	5.40
$SrCO_3$	62		I	4.44	6.78	8.18
SrF_2	225		I	6.74	10.61	11.20
SrLiSb	62		D	0.65	1.09	1.33
SrO	186		D	2.69	5.24	4.47
SrO	216		I	2.58	5.21	4.48
SrO	221		I	2.74	4.04	5.04
SrO	225		I	3.26	4.91	5.57
SrS	186		I	2.80	4.04	4.40
SrS	216		I	3.11	5.06	4.95
SrS	221		I	1.67	2.78	3.41
SrS	225		I	2.41	4.16	4.33
SrSe	186		D	2.40	3.50	3.86
SrSe	216		I	2.78	4.42	4.57
SrSe	221		D	1.29	2.25	2.74
SrSe	225		I	2.13	3.07	3.84
SrTe	186		I	2.05	2.76	3.57
SrTe	216		I	2.84	4.12	4.46
SrTe	221		D	0.36	1.15	1.44
SrTe	225		I	1.64	2.32	3.02
$SrZnF_4$	88		D	5.04	8.28	9.69

TABLE II. (*Continued.*)

Compound	Space group	E_g (exp.; eV)	Type[a]	E_g (PBE; eV)	E_g (mBJ; eV)	$E_g(G_0W_0$; eV)
$Zn(GaTe_2)_2$	82		D	1.02	1.72	2.05
$ZnCl_2$	33		D	4.01	6.32	7.29
ZnF_2	136		D	3.44	6.74	8.09
$ZnGa_2Se_4$	82		D	1.40	2.49	2.80
$ZnGa_2Te_4$	82		D	1.02	1.72	2.05
$ZnIn_2Se_4$	82		D	0.98	1.92	2.19
ZnO	186	3.44[c]	D	0.78	2.63	3.14
ZnO	216		D	0.68	2.52	2.92
ZnO	225		I	0.89	2.89	3.22
ZnS	186		D	2.29	3.88	4.04
ZnS	216	3.66[b]	D	2.22	3.80	3.96
ZnSe	186		D	1.39	2.88	2.99
ZnSe	216	2.70[b]	D	1.34	2.83	2.97
$ZnSiAs_2$	122		D	0.85	1.81	1.74
$ZnSnP_2$	122		D	0.68	1.53	1.67
ZnTe	186		D	1.36	2.50	2.85
ZnTe	216	2.38[b]	D	1.32	2.52	2.77

[a] D and I denote direct and indirect band gaps, respectively.

[b] From Ref. [43].

[c] From Ref. [44].

[d] From Ref. [45].

[1] S. M. Sze and N. N. Ng, *Physics of Semiconductor Devices*, 3rd ed. (John Wiley and Sons, Hoboken, NJ, 2006).

[2] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, Comput. Mater. Sci. **58**, 227 (2012).

[3] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, APL Mater. **1**, 011002 (2013).

[4] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM **65**, 1501 (2013).

[5] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).

[6] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[7] F. Tran and P. Blaha, Phys. Rev. Lett. **102**, 226401 (2009).

[8] M. K. Y. Chan and G. Ceder, Phys. Rev. Lett. **105**, 196403 (2010).

[9] J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. **118**, 8207 (2003); **124**, 219906 (2006).

[10] F. Fuchs, J. Furthmüller, F. Bechstedt, M. Shishkin, and G. Kresse, Phys. Rev. B **76**, 115109 (2007).

[11] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka, Phys. Rev. B **89**, 054303 (2014).

[12] Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, Phys. Rev. B **85**, 104104 (2012).

[13] K. Fujimura, A. Seko, Y. Koyama, A. Kuwabara, I. Kishida, K. Shitara, C. A. J. Fisher, H. Moriwake, and I. Tanaka, Adv. Energy Mater. **3**, 980 (2013).

[14] L. Xu, L. Wencong, P. Chunrong, S. Qiang, and G. Jin, Comput. Mater. Sci. **46**, 860 (2009).

[15] A. Seko, A. Takahashi, and I. Tanaka, Phys. Rev. B **90**, 024101 (2014).

[16] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).

[17] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).

[18] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Phys. Rev. Lett. **114**, 105503 (2015).

[19] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, Int. J. Quantum Chem. **115**, 1094 (2015).

[20] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, Phys. Rev. B **89**, 094104 (2014).

[21] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, arXiv:1508.05315 (2015).

[22] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).

[23] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, Sci. Rep. **3**, 2810 (2013).

[24] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, J. Chem. Theory Comput. **11**, 2087 (2015).

[25] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Muller, and E. K. U. Gross, Phys. Rev. B **89**, 205118 (2014).

[26] W. Setyawan, R. M. Gaume, S. Lam, R. S. Feigelson, and S. Curtarolo, ACS Comb. Sci. **13**, 382 (2011).

[27] P. Dey, J. Bible, S. Datta, S. Broderick, J. Jasinski, M. Sunkara, M. Menon, and K. Rajan, Comput. Mater. Sci. **83**, 185 (2014).

[28] T. Gu, W. Lu, X. Bao, and N. Chen, Solid State Sci. **8**, 129 (2006).

[29] G. Montavon, M. Rupp, V. Gobre, A. Vazquez- Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, New J. Phys. **15**, 095003 (2013).

[30] R. Ramakrishnan, M. Hartmann, E. Tapavicza, and O. A. von Lilienfeld, J. Chem. Phys. **143**, 084111 (2015).

[31] R. Tibshirani, J. R. Stat. Soc.: Ser. B **58**, 267 (1996).

[32] A. J. Smola and B. Schölkopf, Stat. Comput. **14**, 199 (2004).

[33] B. Scholkopf, A. Smola, and K. R. Muller, Neural Comput. **10**, 1299 (1998).

[34] D. Mayer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.-C. Chang, and C.-C. Lin, e1071: Misc Functions of the Department of Statistics, *Probability Theory Group (Formerly:* E1071). TU Wien, R Package version 1.6.-4 (Vienna University of Technology, Vienna, Austria, 2014).

[35] *R Development Core Team.* R*:* A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2015).

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

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

[38] G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996).

[39] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).

[40] 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).

[41] W. M. Haynes, *CRC Handbook of Chemistry and Physics*, 92nd ed. (CRC Press, Boca Raton, FL, 2012).

[42] H. Akaike, in *Second International Symposium on Information Theory*, edited by B. N. Petrov and F. Csaki (Akademiai Kiado, Budapest, Hungary, 1973), pp. 267–281.

[43] J. Heyd, J. E. Peralta, G. E. Scuseria, and R. L. Martin, J. Chem. Phys. **123**, 174101 (2005).

[44] M. Shishkin and G. Kresse, Phys. Rev. B **75**, 235102 (2007).

[45] M. J. Weber, *Handbook of Optical Materials*, 1st ed. (CRC Press, Boca Raton, FL, 2002).