Fundamental band gap and alignment of two-dimensional semiconductors explored by machine learning*

Zhen Zhu(朱震)[1,†], Baojuan Dong(董宝娟)[2,4,5], Huaihong Guo(郭怀红)[3], Teng Yang(杨腾)[2,‡], and Zhidong Zhang(张志东)[2]

[1] *Materials Department, University of California, Santa Barbara, CA* 93106, *USA*
[2] *Shenyang National Laboratory for Materials Science, Institute of Metal Research, Chinese Academy of Sciences, Shenyang* 110016, *China*
[3] *College of Sciences, Liaoning Shihua University, Fushun* 113001, *China*
[4] *State Key Laboratory of Quantum Optics and Quantum Optics Devices, Institute of Opto-Electronics, Shanxi University, Taiyuan* 030006, *China*
[5] *Collaborative Innovation Center of Extreme Optics, Shanxi University, Taiyuan* 030006, *China*

(Received 31 December 2019; revised manuscript received 31 January 2020; accepted manuscript online 13 February 2020)

Two-dimensional (2D) semiconductors isoelectronic to phosphorene have been drawing much attention recently due to their promising applications for next-generation (opt)electronics. This family of 2D materials contains more than 400 members, including (a) elemental group-V materials, (b) binary III–VII and IV–VI compounds, (c) ternary III–VI–VII and IV–V–VII compounds, making materials design with targeted functionality unprecedentedly rich and extremely challenging. To shed light on rational functionality design with this family of materials, we systemically explore their fundamental band gaps and alignments using hybrid density functional theory (DFT) in combination with machine learning. First, calculations are performed using both the Perdew–Burke–Ernzerhof exchange–correlation functional within the general-gradient-density approximation (GGA-PBE) and Heyd–Scuseria–Ernzerhof hybrid functional (HSE) as a reference. We find this family of materials share similar crystalline structures, but possess largely distributed band-gap values ranging approximately from 0 eV to 8 eV. Then, we apply machine learning methods, including linear regression (LR), random forest regression (RFR), and support vector machine regression (SVR), to build models for the prediction of electronic properties. Among these models, SVR is found to have the best performance, yielding the root mean square error (RMSE) less than 0.15 eV for the predicted band gaps, valence-band maximums (VBMs), and conduction-band minimums (CBMs) when both PBE results and elemental information are used as features. Thus, we demonstrate that the machine learning models are universally suitable for screening 2D isoelectronic systems with targeted functionality, and especially valuable for the design of alloys and heterogeneous systems.

Keywords: two-dimensional semiconductors, machine learning

PACS: 73.61.Cw, 61.46.–w, 73.22.–f **DOI:** 10.1088/1674-1056/ab75d5

1. Introduction

Last decade has witnessed the rocketing development of two-dimensional (2D) materials, which find promising applications in next-generation electronics and optoelectronics.[1–4] The performance of a 2D electronic device depends sensitively on the fundamental electronic properties of the candidate material: a non-zero band gap, proper band edge positions, and high mobility are in general the requisites. In contrary to semi-metallic graphene[1,2] and low-mobility transition metal dichalcogenides (TMDs)[5,6] that fail to deliver good device performance, phosphorene is semiconducting while still maintaining a high hole mobility,[7–11] thereby emerging as a potential candidate for 2D electronics. However, poor chemical stability has limited its practical applications.[12] To overcome such obstacles, searching for 2D materials with similar electronic properties but better chemical stability is essential.

Recently, high-throughput materials screening has emerged as an effective method to search for materials with targeted functionality.[13–16] The workflow for materials discovery is separated to different layers: starting with crude and low-precision computations to narrow the candidacy pool, and followed by precise but expensive calculations to identify the candidate materials. The initial materials pool is usually a subset of the ICSD database[17] with large amount of candidates, resulting in tedious prescreening and large computational efforts. A prescreening method that is both accurate and computationally efficient is greatly desired, where machine learning can play an important role. In combination with density functional theory (DFT), machine learning has demonstrated valuable applications in functional materials design,[18] properties predictions,[19–22] and many other fields[23,24] for traditional bulk materials. It is intriguing to apply such machine learning methods to two-dimensional systems to accelerate materials discovery, which is largely unexplored but fundamentally and technologically important.

*This work is dedicated to Michelle Mucheng Zhu. Project supported by the National Key R&D Program of China (Grant No. 2017YFA0206301).
†Corresponding author. E-mail: zhuzhen@engineering.ucsb.edu
‡Corresponding author. E-mail: yangteng@imr.ac.cn

Here, we have explored the fundamental band gaps and band alignments of a group of 2D semiconductors that are isoelectronic to phosphorene using machine learning techniques in combination with density functional theory. The methodology is discussed in Section 2, including details of density functional calculations (Subsection 2.1) and a brief introduction of machine learning models (Subsection 2.2). We describe the isoelectronic materials design method in Subsection 2.3. Following this method, more than 400 materials are constructed and calculated, including (a) elemental group-V materials, (b) binary III–VII and IV–VI compounds, and (c) ternary III–VI–VII and IV–V–VII compounds. Among this family of materials, many have been successfully synthesized[25,26] and found special applications in different research fields.[27,28] The richness in electronic properties of these materials is categorized and analyzed in Section 3. Next, in Section 4, we apply machine learning methods, including linear regression (LR), random forest regression (RFR), and support vector machine regression (SVR), to predict electronic properties for this family of 2D materials. Then we summarize our key findings in Section 5.

2. Methodology

2.1. Computational details of density functional methods

All our calculations are based on DFT using projector-augmented waves[29] (PAW) as implemented in the VASP[30] code. We have used periodic boundary conditions throughout the study, with monolayer structures represented by a periodic array of slabs separated by a vacuum region at least 15 Å thick. We use the Perdew–Burke–Ernzerhof (PBE)[31] exchange–correlation functional for the initial structure optimization based on the conjugate gradient method[32] with a 400 eV energy cutoff. All geometries are treated as optimized when none of the residual Hellmann–Feynman forces exceed 10^{-2} eV/Å. On top of PBE-optimized structures, a single-shot screened hybrid functional calculation (HSE)[33,34] is performed to obtain the fundamental band gap and alignment of the material. We have used standard values for the mixing parameter (0.25) and the range-separation parameter (0.2 $Å^{-1}$). The reciprocal space is sampled by a grid[35] finer than $10\times10\times1$ k-points in the Brillouin zone of the primitive unit cell.

2.2. Machine learning methods

The obtained DFT results are then analyzed with machine learning models as implemented in scikit-learn[36] package. The relation between target electronic properties and predictors can be established via supervised learning methods. A good predictive model depends sensitively on the choice of regression models, selection of predictors, as well as the quality of our dataset. For a given data set, it is important to select proper predictors and suitable regression models to achieve good predictive ability with high accuracy. To achieve this goal in current study, we have selected three different predictor sets, which are different combinations of the computed PBE results and fundamental signatures of constituent elements. Then, we utilize a variety of regression methods, including linear regressions, random forest regression, and support vector machine regression, to predict the target electronic properties.

In the LR method, the regression coefficients of predictors, w, are determined by optimizing the following cost function $L(\boldsymbol{w})$: $L(\boldsymbol{w}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2$. In addition, other LR methods with regularizations, LASSO and Ridge, are also used in this study. On top of the ordinary least square linear regression method, LASSO includes an additional $L1$ penalty term $\sum_i ||\alpha w_i||$ in the cost function, while the Ridge regression method adds an $L2$ regularization term $\sum_i ||\alpha w_i||^2$. These penalty terms can effectively mitigate the overfitting problem especially when the predictor sets are large.

When the relation between the target property and the predictors is not linear, regression methods like RFR and SVR with a non-linear kernel are supposed to capture the nonlinear feature–target relationship. Random forest is one type of ensemble methods. It grows a number of decision trees via bootstrapping the sample space. For each decision tree, a randomly selected subset of the feature space is used, which can effectively minimize the correlation between different trees. Then, the target value is predicted by majority vote of these trees for classification or averaging the predicted result of each tree in regression problems. Importantly, the random forest model is easy to interpret and it can output the relative importance of different features, thereby providing insights on the elemental signatures that determine the targeted electronic properties of materials in the present study.

We also use a SVR model with a radial basis function (RBF) kernel to predict the calculated electronic properties with fundamental materials features. The support vector machine model utilizes the kernel trick to map low-dimensional non-separable data to a higher dimension where they can be separated via a hyper-plane. The optimized hyper-plane can be identified by the so-called supported vectors. The kernel trick makes it possible to compute the inner product of the projected data in the higher dimension without specifying the mapping function, which is usually time-consuming or even impossible to specify. SVR uses a hinge-loss function $\sum_i \max(0, 1 - y_i f(x_i))$, which is minimized during the model training process. The RBF kernel used in the present work has the form of $K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2)$.

Table 1. Three sets of predictors used for machine learning models to predict electronic band gaps and band alignments.

Target property	Predictors set-I	Predictors set-II	Predictors set-III
E_g (HSE)	E_g (PBE)	elements signatures	E_g (PBE), elements signatures
VBM (HSE)	VBM (PBE)	elements signatures	VBM (PBE), elements signatures
CBM (HSE)	CBM (PBE)	elements signatures	CBM (PBE), elements signatures

In addition to the type of machine learning methods we choose, a proper selection of feature space is also critical to achieve robust and accurate prediction. Previous studies usually include a large amount of predictors in the feature space and then conduct dimension reduction, which is likely to hide important physical insights of the model. Here, instead, we intend to compare the prediction power of PBE results as features and merely fundamental chemical and physical signatures of constituent elements in the materials. With this consideration, as shown in Table 1, we have built three different sets of predictors: In set-I, PBE results, band gap, valence-band maximum (VBM), and conduction-band minimum (CBM) are the only features used to predict related HSE values; In set-II, we include only elemental signatures for each material, such as atomic mass, ionization energy, electron affinity, electronegativity, as well as electronegativity difference between cations and anions; set-III is a combination of set-I and set-II. The features in set-I depend on less time-consuming PBE calculations, while set-II is more convenient to obtain with no requirement of any DFT calculations.

2.3. Isoelectronic materials design

The 2D group-V elemental materials, such as phosphorene[8,37] and antimonene,[38] can be stabilized in two distinct structural phases, the black and blue phosphorene phases, as defined to be phases I and II in Fig. 1(a) and 1(b), respectively. In both structural phases, each atom forms three covalent bonds of sp^3 type with adjacent atoms, as well as lone-pair electrons, which fulfils the octet rule. The pyramid formed by a center atom and its three nearest neighbors can be arranged in a variety of ways, thereby leading to a rich design space for structural polymorphs.[39]

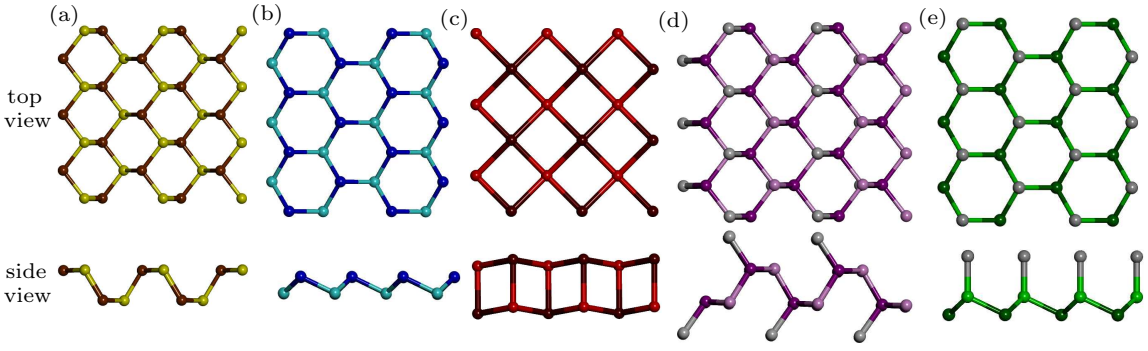

Fig. 1. Equilibrium structures of elemental materials and binary compounds in black-phosphorene-type phase-I, (b) blue-phosphorene-type phase-II, (c) indium-iodide-type phase-III, and their ternary counterparts in (d) phase-I and (e) phase-II. All structure are shown in both top and side views. The different colors represent the V-group, or IV–VI, or III–VII elements without any specifications, except that the brown in (c) represents the III-group element, the dark purple in (d) and the light green in (e) represent the III- or IV-group element, and the grey in (d), (e) represents the VII-group element.

Binary compounds can be derived from their elemental counterparts by cation mutation while the averaged valence electrons are conserved to be five.[40] Based on such a principle, group IV–VI and III–VII compounds can be conveniently designed and they are isoelectronic to the well-studied group-V elemental materials. In addition to two base structures mentioned previously, III–VII compounds can also be stabilized in a special structure with a primitive cell of approximately square shape, as defined to be the phase III in Fig. 1(c). In fact, for indium iodide, an existing compound of the III–VII family, phase-III is the most energetically favored structure among the polymorphs mentioned here.[41,42] Thus, we also include this structural phase as one of the base structures for the 2D materials design in the present study.

The isoelectronic design principle can be further generalized to construct ternary compounds. As shown in Figs. 1(d) and 1(e), III–VI–VII and IV–V–VII compounds share similar structures with the elemental and binary materials discussed above; indeed, they are isoelectronic. Taking phosphorene of phase-II as the starting material, we change half of the P atoms to a group IV element, such as Si. The sp^3 bonding in the material is maintained and the P atoms still have the close-shell electron configuration. However, as Si has one less valence electron, one unpaired electron exists for Si rather than a lone pair in P. Furthermore, a group-VII halogen element can form an additional bond with Si, thereby satisfying the octet rule for the ternary compounds. Therefore, the IV–V–VII compounds are isoelectronic to the group-V elemental materials. Simi-

larly, the III–VI–VII compounds can be shown as isoelectronic counterparts to the IV–VI compounds. Importantly, in the element mutation process to construct the ternary compounds, half of the lone pairs in the original materials no longer exist, but instead form covalent bonds between the metal and halogen atoms. In fact, ternary compounds are not limited to these two groups of materials. Simply applying the cation-mutation principle to the IV–VI and III–VII compounds, we can obtain III–V–VI_2 and II–IV–VII_2 ternary compounds. As they are expected to be rather similar to their parent binary compounds, these groups of ternary materials are not computed using DFT methods in the present work, but instead their electronic properties can be predicted from our machine learning models that are to be discussed in Subsection 4.4.

To build a database for this family of 2D materials, we have considered entire group-III, group-IV, group-V, group-VI, and group-VII elements (except the radiative Tl, Po, and At) for isoelectronic materials design. For elemental and binary materials, three structural phases, phase-I, phase-II, and phase-III, are treated as the base structures to perform element mutation. Following the design principle above, we have constructed 15 elemental materials and 108 binary compounds. For ternary compounds, we only use phase-I and phase-II as the base to construct isoelectronic compounds, giving 328 distinct 2D materials. Then, we perform DFT calculations to obtain the optimized structures, the fundamental band gaps, and the absolute positions of band edges at both PBE and HSE levels. In fact, not all the element combinations can maintain the structural phases we are interested in the present work. Especially, materials containing B, C, O, and N are in general not able to be stabilized in the desired structural form. The data points corresponding to these materials are eliminated from the database and not used for machine learning exploration.

3. Electronic properties by DFT-PBE and HSE

The calculated electronic properties, fundamental band gaps, VBMs, and CBMs, are shown in Fig. 2. It is known that for the same semiconductor, the HSE band gap would scale linearly with the mixing parameter and the DFT-PBE band gap value is in general the intercept. However, for different materials, it is not clear how the HSE band gaps are related to that predicted by DFT-PBE. Here in Fig. 2, we illustrate that the HSE band-gap values scale approximately linearly with that of DFT-PBE. The linear relation can be further improved when these isoelectronic materials are separated to different categories based on the number of constituent elements, which is reflected by the color-distinguished data points in Fig. 2(a).

For the absolute positions of band edges, the linear relationship between HSE and DFT-PBE is even more clear. The VBM position of a material, referenced to the vacuum level, corresponds to its electron ionization energy, which in general can be predicted by HSE to a good agreement with experiments. As shown in Fig. 2(b), HSE predicts lower VBMs than those of PBE and we also find a linear relationship between VBMs of HSE and PBE. Therefore, promisingly, the PBE results may act as efficient descriptors for expensive HSE calculations, as well as experimental results, which is to be assessed in Section 4. Similarly, HSE CBMs also scale linearly with that obtained by PBE, illustrated in Fig. 2(c). However, for CBMs, the HSE results are slightly higher than those of PBE, in sharp contrast to the case for VBMs.

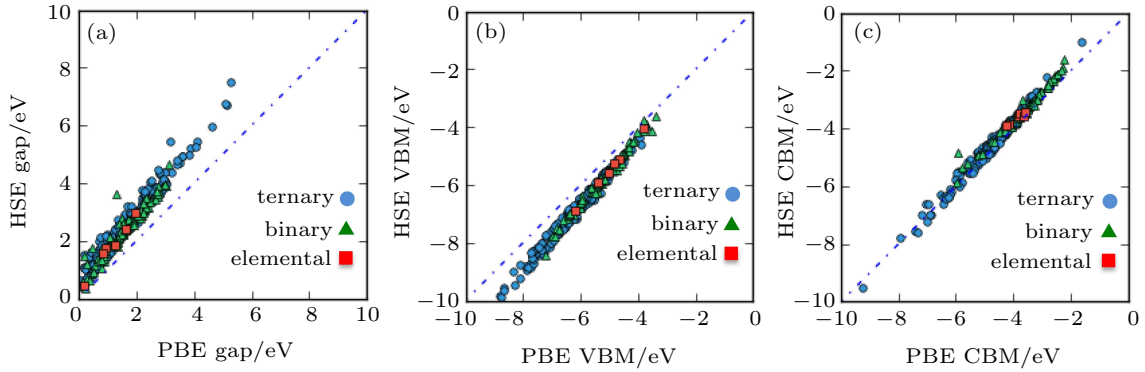

Fig. 2. Relation between PBE and HSE results of (a) fundamental band gaps, (b) VBMs, and (c) CBMs. The dashed lines are guide to the eyes, indicating the case of HSE values equaling PBE values.

To gain deeper insights into the electronic properties of this family of materials, we have shown the distributions of band gaps, VBMs, and CBMs with respect to both materials types and structural phases (Fig. 3). Clearly, for both elemental and binary materials, the phase-II structures have larger band gap values than those of phase-I, which is closely related to the fact that VBMs of the former are in general lower than those of the later, indicated in Fig. 3(b). The similar trend for these two groups of materials can be explained by the fact that they are isoelectronic and the band-edge states are similar. On the contrary, for the ternary compounds, the averaged band gap value of the phase-I structures is larger than that of the phase-II structures by ~ 1 eV. Especially, they have very similar distributions of VBMs. The different behaviors between

the ternary compounds and the others is due to the presence of halogen ligand that satisfies the octet rule without forming lone-pair electrons. In fact, the ternary compounds are not "perfectly" isoelectronic to their parent compounds. Since the signature of long-pair electrons is still partially persevered in the ternary compounds, CBMs share similar characters for all three types of materials (Fig. 3(c)). Furthermore, the binary compounds are found to have larger averaged band gap than that of elemental materials, which can be attributed to the increased ionicity in the materials: a larger electronegativity difference between cation and anion usually leads to a larger band gap value.[43] The factors that affect the band gaps and alignments of materials are to be discussed in Subsection 4.4.

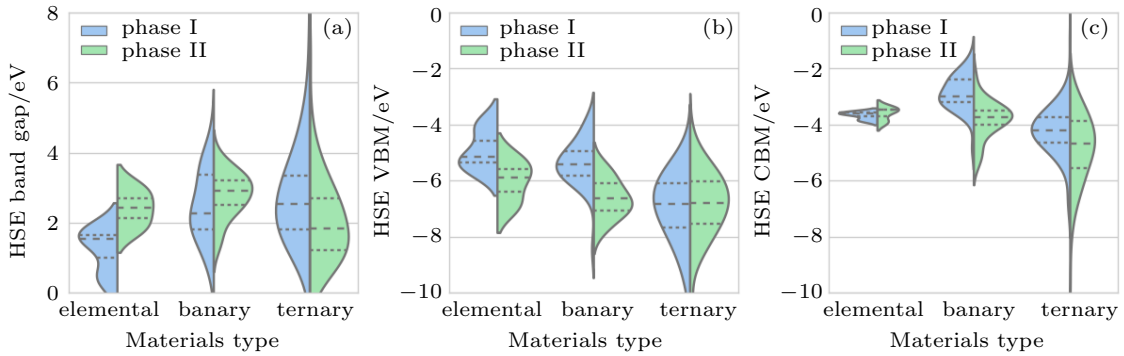

Fig. 3. Distributions of (a) band gap values, (b) VBMs, and (c) CBMs based on HSE calculations for different materials types. The blue-colored areas represent materials of phase-I, while the green-colored ones are for phase-II. The long-dashed line indicates the mean of the distribution.

4. Machine learning predictive models

As mentioned in Section 2, computational results in the present work are at two distinct levels of theory: DFT-PBE and the screened hybrid functional method (HSE). The former is computationally less demanding, but severely underestimates the fundamental band gap; HSE, on the other hand, can precisely predict the band-gap values and alignments of standard semiconductors (without localized d or f orbitals as valence electrons), but is formidable for large-scale functional materials screening due to high computational cost. Therefore, it is desirable to build computational efficient methods that can also achieve high accuracy simultaneously. Given different predictor sets as described in Subsection 2.2, we apply machine learning methods, including LR, RFR, and SVR, to predict computed electronic properties at HSE level, which can be further utilized to predict experimental observations.

4.1. Set-I predictors

As inferred from Section 3, the HSE band gaps have approximately linear relation with that of PBE. Intuitively, the PBE band gaps have been used as the only feature in predictors set-I. The LR model is applied to model the relation between the results of HSE and PBE with 10-fold cross validation, while RFR and SVR are not applicable for such simple feature space. The predicted HSE gaps of the validation sets are shown with respect to the calculated values in Fig. 4(a). The relation is

$$E_{\mathrm{g}}^{\mathrm{HSE}} = 1.21E_{\mathrm{g}}^{\mathrm{PBE}} + 0.52\ \mathrm{eV}. \quad (1)$$

The residues, difference between the predicted and computed band-gap values, are presented in Fig. 4(d) and the majority fall into the $[-0.5\ \mathrm{eV}, 0.5\ \mathrm{eV}]$ energy range, indicating good prediction accuracy. There are only two data points where the difference is larger than 1.0 eV. Even though they might be outliers, their influence on our regression model is minimal. Furthermore, we also calculate the root mean square error (RMSE) and the mean absolute percent error (MAPE) to evaluate the predictive model and the small prediction error, 0.25 eV for RMSE and 10.67% for MAPE (see Table. 2), also reflects the high accuracy of the model.

Similarly, in order to predict $\mathrm{VBM}^{\mathrm{HSE}}$ positions, the $\mathrm{VBM}^{\mathrm{PBE}}$ values are used as the only feature in set-I predictors space. The predicted $\mathrm{VBM}^{\mathrm{HSE}}$ positions of validation sets are presented in Fig. 4(b), showing excellent agreement with targeted values. All the residues [Fig. 4(e)] are in the $[-0.5\ \mathrm{eV}, 0.5\ \mathrm{eV}]$ energy range. The better linearity of the VBM predictive model, comparing with that of band gap, also leads to smaller prediction errors as listed in Table. 2. The predicted relationship between $\mathrm{VBM}^{\mathrm{PBE}}$ and $\mathrm{VBM}^{\mathrm{HSE}}$ is

$$\mathrm{VBM}^{\mathrm{HSE}} = 1.15\mathrm{VBM}^{\mathrm{PBE}} + 0.23\ \mathrm{eV}. \quad (2)$$

$\mathrm{CBM}^{\mathrm{HSE}}$ can also be predicted by $\mathrm{CBM}^{\mathrm{PBE}}$ with LR method and the model accuracy is illustrated in Figs. 4(c) and 4(f). For CBM, the relation between the HSE and PBE results is

$$\mathrm{CBM}^{\mathrm{HSE}} = 1.07\mathrm{CBM}^{\mathrm{PBE}} + 0.51\ \mathrm{eV}. \quad (3)$$

Since these three linear models are cross-validated by randomly selected samples from our 2D materials data set, they should be universally valid for materials that are isoelectronic to current family members.

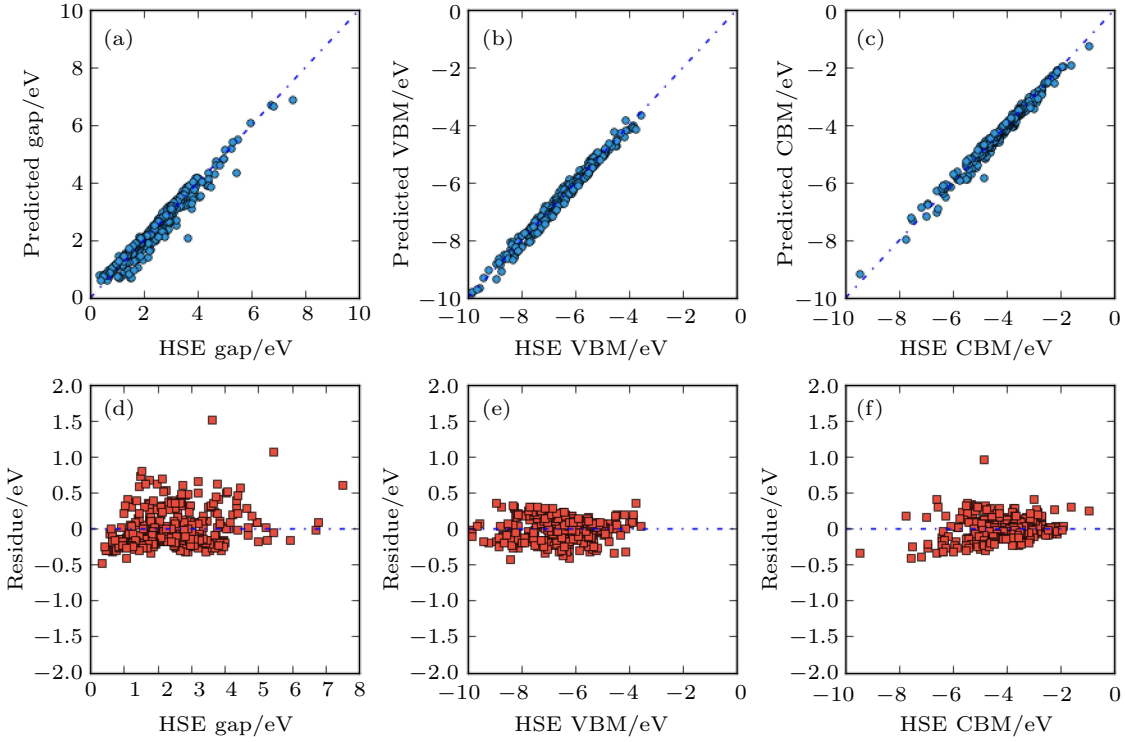

Fig. 4. Comparison of predicted (a) band gaps, (b) VBMs, and (c) CBMs by LR with calculated HSE values. Computed relevant PBE value (predictors set-I) is used as the single descriptor for the predictive LR model. Corresponding residues are shown in (d), (e), and (f) to access the accuracy of the model. The dashed lines are guide to the eyes, representing that the predicted values are equal to the computed HSE data.

Table 2. Prediction errors of band gaps in the LR, RFR, and SVR models.

Regr. methods	Pred. sets	Band gap RMSE/eV	Band gap MAPE/%	VBM RMSE/eV	VBM MAPE/%	CBM RMSE/eV	CBM MAPE/%
LR	set-I	0.25	10.67	0.15	1.85	0.14	2.53
	set-II	0.87	35.07	0.88	10.30	0.80	16.03
	set-III	0.15	5.55	0.09	1.04	0.09	1.56
RFR	set-I	–	–	–	–	–	–
	set-II	0.70	26.37	0.67	7.23	0.57	10.22
	set-III	0.25	7.44	0.18	1.75	0.18	2.64
SVR	set-I	–	–	–	–	–	–
	set-II	0.57	16.80	0.49	4.83	0.43	7.07
	set-III	0.13	4.93	0.08	0.96	0.09	1.65

4.2. Set-II predictors

The ideal predictive model would rather have elemental information of constituent elements as the feature space, instead of DFT results at any level of theory. This would greatly improve the model efficiency and even make real-time interactive prediction possible. We have created predictors set-II to fulfill such a purpose. Details about this set of predictors are discussed in Subsection 2.2.

For this set of predictors, we have applied LR, RFR, and SVR to predict the targeted electronic properties. The performances of these models are shown in Fig. 5, as inferred from the relationship between the predicted values and computed values. Here the LR model shows much inferior predictive ability comparing with the case when the PBE results are used as predictors. This is also reflected by its high RMSE (0.87 eV) and high MAPE (35.07%) as presented in Table 2. Comparing with band-gap prediction, the accuracy of the LR model is slightly improved for VBMs and CBMs: MAPE values are 10.30% and 16.03%, respectively. To avoid overfitting, we have also compared the simple LR model with regularized models, such as Ridge regression and LASSO, and found no improvement in the model performance.

The undesired performance of the LR model indicates that the nonlinear relationship between the set-II predictors and computed HSE results is essential. Complicated models, like RFR and SVR, are likely to capture the nonlinearity in the feature–target relation. Indeed, we find both RFR and SVR models have better performance than the former LR model. SVR is found to give the lowest RMSEs: 0.57 eV

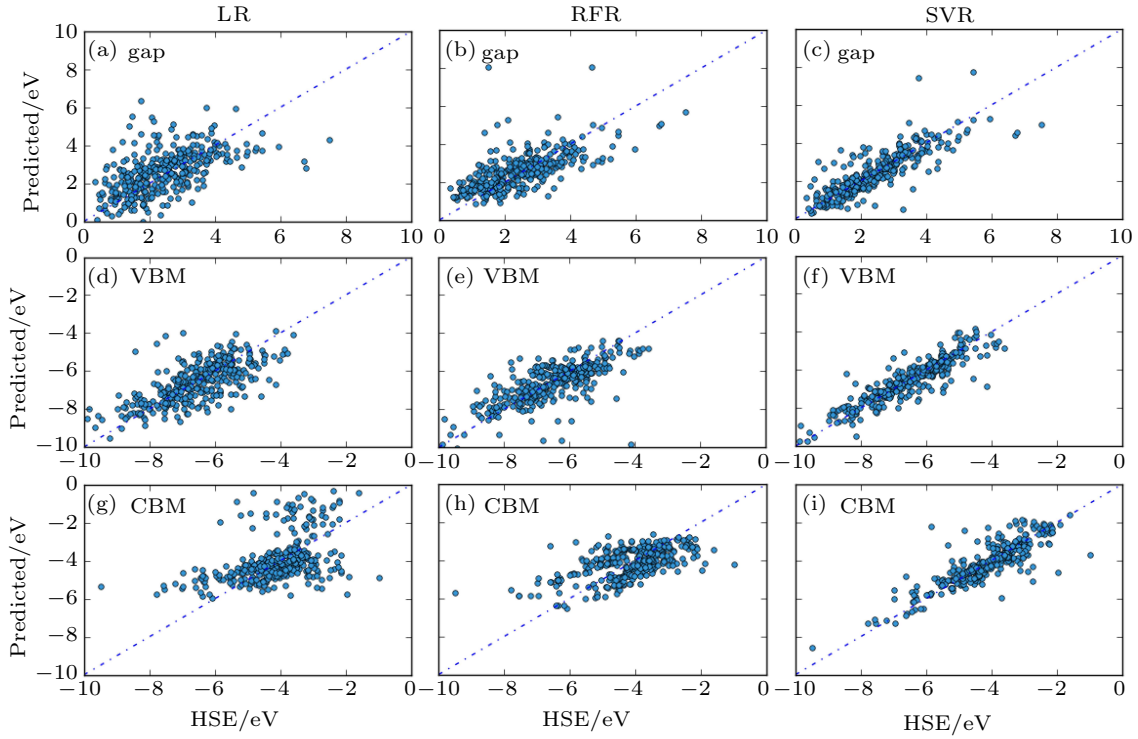

Fig. 5. Comparison of calculated HSE values with predicted fundamental band gaps and band edges by LR, RFR, and SVR models. These models are evaluated with 10-fold cross-validation and only the predicted results of the validation sets are shown. Each column of the subfigures represents one predictive model and each row shows one selected electronic property. Selected signatures of constituent elements are used as predictors (predictors set-II) for these machine learning models. The dashed lines are guide to the eyes, representing that the predicted values are equal to the computed HSE data.

for band gaps, 0.49 eV for VBMs, and 0.43 eV for CBMs, corresponding to MAPEs of 16.80%, 4.83%, and 7.07%, which achieve approximately 50% error reduction from the LR model. Even though the performance is still inferior to LR with DFT-PBE results as the features, it should be noted that the SVR model we developed here is of advantage to be used for fast materials screening due to its convenient feature space with no requirements for DFT calculations.

Although RFR is not the best predictive model, it can provide precious insights into important features that determine the underlying materials properties. Alongside training of a RFR model, we can also obtain the relative importance of predictors in the feature space. For band-gap prediction, the most significant feature is the average mass: the heavier the compounds, the smaller the band gap. It is noted that increased metallicity is inherited naturally from larger atomic mass for elements from the same element group, which weakens both bonding strength and ionicity, resulting in the narrowing of the band gap. Other important features include the electronegativity difference between cation and anion, cation electronegativity, phase type, and so on. For VBMs and CBMs, the rankings of feature importance are different: the average mass is not as important as for the band-gap prediction. VBMs depend strongly on the electronegativity difference between cation and anion, while the anion electron affinity is the most significant factor determining CBMs.

4.3. Set-III predictors

The predictive models can be further improved when predictors of set-I and set-II are combined as the new feature space: set-III predictors. As DFT-PBE can also be viewed as a good predictive model, machine learning methods based on set-III predictors can thus be viewed as a process of model stacking, which in general give better prediction performance. The predicted results for the validation sets are compared with the computed values in Fig. 6. Indeed, we find that the RMSE and MAPE of all three regression models are significantly reduced with respect to the case where the feature space is spanned by either set-I or set-II predictors. Among the machine learning models used here, SVR outperforms the other two for all three targeted materials properties, with RMSEs of 0.13 eV for band gap, 0.08 eV for VBM, and 0.09 eV for CBM. In fact, the prediction errors are within the accuracy of HSE calculations, thereby justifying the validity and accuracy of our models for properties prediction.

4.4. Discussion

Model selection By carefully comparing the performance of different machine learning models, we have elucidate the general principle for model selection. Both RMSE and MAPE are computed to evaluate the accuracy of LR, RFR, and SVR models. Among these three models in the present study, SVR is found to have the best accuracy, when either set-II or set-III predictors are used as the feature space. Especially,

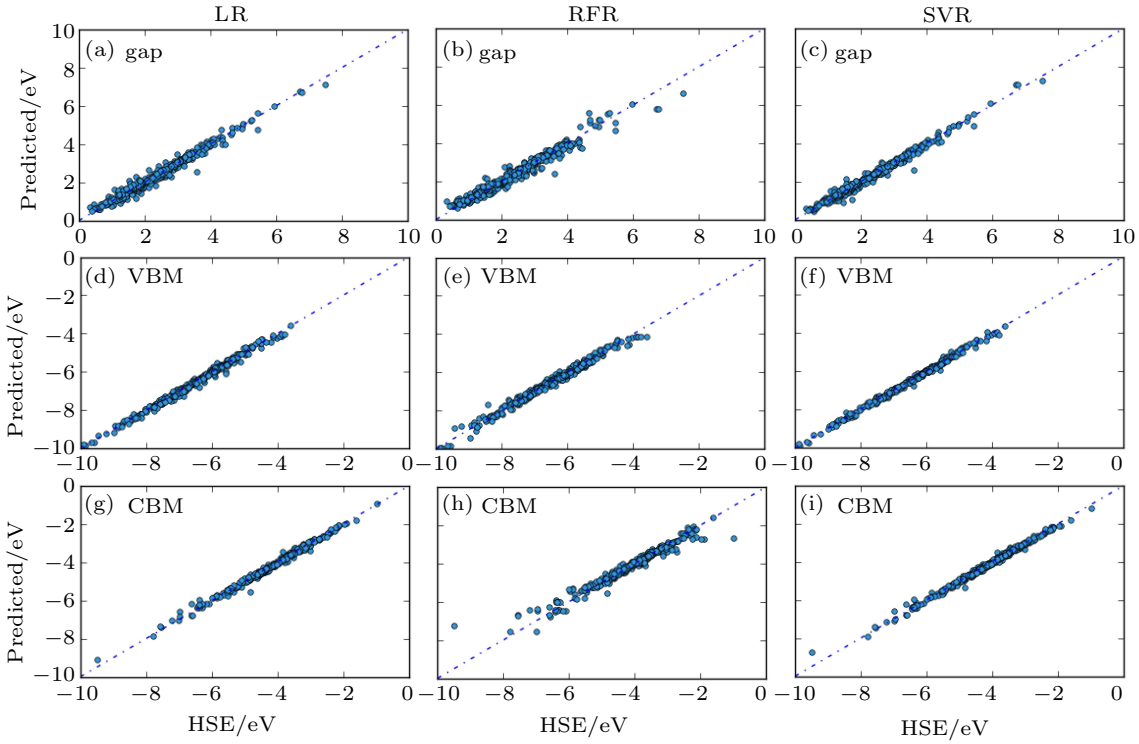

Fig. 6. Comparison of calculated HSE values with predicted fundamental band gaps and band alignments by LR, RFR, and SVR models, where both relevant PBE results and selected signatures of constituent elements are used as predictors (predictors set-III). These models are evaluated with 10-fold cross-validation and only the predicted results of the validation sets are shown. Each column of the subfigures represents one predictive model and each row shows one selected electronic property. The dashed lines are guide to the eyes, representing that the predicted values are equal to the computed HSE data.

when only elemental information is used to span the feature space, SVR shows significant advantage in predicting targeted materials properties over the other two methods. Therefore, SVR is suggested to use when no prior DFT-PBE results are available. On the other hand, if DFT-PBE values are available, LR is a good model to start with. In this method, the relation between HSE values and DFT-PBE features can be expressed in a simple analytical model, thus the target values can be readily predicted.

Performance for different target properties The same machine learning model is found to have different performances when target properties vary. Even though the band gap is closely related to VBM and CBM, the later two targets almost always have smaller RMSE and MAPE than the former. The difference in accuracy is likely caused by the fact that for band-gap prediction, a good predictor reflecting both VBM and CBM states is a requisite, which is unlikely to be included in our simple feature space. On the other hand, for VBM- or CBM-prediction, the requirement is less stringent and more likely to be covered by our selection of predictors. To further improve the model performance, we expect to have a more complicated feature space, including different operations between predictors in current feature space. This is beyond the scope of current study.

Applications As mentioned in Subsection 2.3, the materials used in the present work are just a small fraction of this large family of materials following proposed isoelectronic materials design principle. Our trained models, especially SVR with set-II predictors, can be applied to predict the fundamental band gaps and alignments of other family members with minimal computation cost. The predicted results are informative and valuable even when the designed materials are not the most stable structural phase. It has been shown that alloying unstable materials with stable ones in the desired structural phase is likely to stabilize the former compounds. For example, CaSe can be stabilized in phase-I when alloying with SnSe.[44] The electronic properties of such alloys can also be predicted by our models where the weighted average of the constituent elements are taken as predictors. Therefore, the trained machine learning models in the present study provide a computational efficient method to accurately obtain the band gaps and alignments of a large amount of 2D materials, which enables fast screening of 2D functional materials for electronic, optoelectronic, and photocatalysis applications.

5. Conclusions

We have explored fundamental band gaps and alignments of a group of two-dimensional semiconductors isoelectronic to phosphorene using machine learning techniques in combination with density functional theory. This family of 2D materials shares similar crystalline structures, but possesses unprecedented rich band-gap values ranging approximately from 0 eV to 8 eV. Based on the machine learning methods, we

trained predictive models that can predict band-gap values and band-edge positions with surprisingly high accuracy. Among models discussed in the present work, SVR is found to have the best performance with RMSEs less than 0.15 eV for the predicted band gaps, VBMs, and CBMs when both PBE results and elemental information are used as predictors. We also demonstrate that the predictive models can be utilized for electronic properties prediction for more complicated systems, like quaternary compounds and alloys, shedding light on rational materials design for (opto)electronic and photocatalysis applications.

References

[1] Novoselov K S, Geim A K, Morozov S V, Jiang D, Zhang Y, Dubonos S V, Grigorieva I V and Firsov A A 2004 *Science* **306** 666
[2] Zhang Y, Tan Y W, Stormer H L and Kim P 2005 *Nature* **438** 201
[3] Mak K F and Shan J 2016 *Nat. Photon.* **10** 216
[4] Fiori G, Bonaccorso F, Iannaccone G, Palacios T, Neumaier D, Seabaugh A, Banerjee S K and Colombo L 2014 *Nat. Nanotechnol.* **9** 768
[5] Radisavljevic B, Radenovic A, Brivio J, Giacometti I V and Kis A 2011 *Nat. Nanotechnol.* **6** 147
[6] Perera M M, Lin M W, Chuang H J, Chamlagain B P, Wang C, Tan X, Cheng M M C, Tománek D and Zhou Z 2013 *ACS Nano* **7** 4449
[7] Li L, Yu Y, Ye G J, Ge Q, Ou X, Wu H, Feng D, Chen X H and Zhang Y 2014 *Nat. Nanotechnol.* **9** 372
[8] Liu H, Neal A T, Zhu Z, Luo Z, Xu X, Tománek D and Ye P D 2014 *ACS Nano* **8** 4033
[9] Koenig S P, Doganov R A, Schmidt H, Castro Neto A H and Ozyilmaz B 2014 *Appl. Phys. Lett.* **104** 103106
[10] Rodin A S, Carvalho A and Castro Neto A H 2014 *Phys. Rev. Lett.* **112** 176801
[11] Ling X, Wang H, Huang S, Xia F and Dresselhaus M S 2015 *Proc. Natl. Acad. Sci. USA* **112** 4523
[12] Liu H, Du Y, Deng Y and Peide D Y 2015 *Chem. Soc. Rev.* **44** 2732
[13] Curtarolo S, Hart G L, Nardelli M B, Mingo N, Sanvito S and Levy O 2013 *Nat. Mater.* **12** 191
[14] De Jong M, Chen W, Angsten T, *et al.* 2015 *Scientific Data* **2** 150009
[15] Setyawan W and Curtarolo S 2010 *Computational Materials Science* **49** 299
[16] Hautier G, Jain A, Ong S P, Kang B, Moore C, Doe R and Ceder G 2011 *Chemistry of Materials* **23** 3495
[17] https://icsd.fiz-karlsruhe.de
[18] Ward L and Wolverton C 2017 *Current Opinion in Solid State and Materials Science* **21** 167
[19] Ward L, Agrawal A, Choudhary A and Wolverton C 2016 *npj Computational Materials* **2** 16028
[20] Deml A M, O'Hayre R, Wolverton C and Stevanović V 2016 *Phys. Rev. B* **93** 085142
[21] Pilania G, Gubernatis J E and Lookman T 2017 *Computational Materials Science* **129** 156
[22] Lee J, Seko A, Shitara K, Nakayama K and Tanaka I 2016 *Phys. Rev. B* **93** 115104
[23] Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A and Kim C 2017 *npj Computational Materials* **3** 54
[24] Balachandran P V, Theiler J, Rondinelli J M and Lookman T 2015 *Sci. Rep.* **5** 13285
[25] Ji J, Song X, Liu J, *et al.* 2016 *Nat. Commun.* **7** 13352
[26] Zhang J L, Zhao S, Han C, *et al.* 2016 *Nano Lett.* **16** 4903
[27] Haleoot R, Paillard C, Mehboudi M, Xu B, Bellaiche L and Barraza-Lopez S 2017 *Phys. Rev. Lett.* **118** 227401
[28] Fei R, Kang W and Yang L 2016 *Phys. Rev. Lett.* **117** 097601
[29] Blöchl P 1994 *Phys. Rev. B* **50** 17953
[30] Kresse G and Furthmüller J 1996 *Phys. Rev. B* **54** 11169
[31] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865
[32] Hestenes M R and Stiefel E 1952 *J. Res. Natl. Bur. Stand.* **49** 409
[33] Heyd J, Scuseria G E and Ernzerhof M 2003 *J. Chem. Phys.* **118** 8207
[34] Heyd J, Scuseria G E and Ernzerhof M 2003 *J. Chem. Phys.* **124** 219906
[35] Monkhorst H J and Pack J D 1976 *Phys. Rev. B* **13** 5188
[36] https://http://scikit-learn.org/
[37] Zhu Z and Tománek D 2014 *Phys. Rev. Lett.* **112** 176802
[38] Zhang S, Yan Z, Li Y, Chen Z and Zeng H 2015 *Angewandte Chemie* **127** 3155
[39] Guan J, Zhu Z and Tománek D 2014 *ACS Nano* **8** 12763
[40] Zhu Z, Guan J, Liu D and Tománek D 2015 *ACS Nano* **9** 8284
[41] Wang J, Dong B J, Guo H, Yang T, Zhu Z, Hu G, Saito R and Zhang Z D 2017 *Phys. Rev. B* **95** 045404
[42] Zhang Y, Guo H, Dong B J, Zhu Z, Yang T, Wang J and Zhang Z 2020 *Chin. Phys. B* **29** 037305
[43] Goodman C 1958 *J. Phys. Chem. Solids* **6** 305
[44] Matthews B E, Holder A M, Schelhas L, *et al.* 2017 *J. Mater. Chem. A* **5** 16873