ML-aided computational screening of 2D materials for photocatalytic water splitting

Yatong Wang,[1,2,‡] Murat Cihan Sorkun,[1,‡] Geert Brocks,[2,3] and Süleyman Er[1,*]

ABSTRACT

The exploration of two-dimensional (2D) materials with exceptional physical and chemical properties is essent ial for the advancement of solar water-splitting technologies. However, the discovery of 2D materials is currently heavily reliant on fragmented studies with limited opportunities for fine-tuning the chemical composition and electronic features of compounds. Here, we apply a combination of machine learning (ML) and physics-based computation to evaluate the V2DB digital library, which contains an extensive collection of 2D materials for their potential use in photocatalytic water splitting. To examine the structural and electronic properties of the potential 2D photocatalysts, we utilize a computational funnel approach that integrates ML modeling, as well as DFT, hybrid-DFT, and GW calculations. Our screening process yields a selection of 11 promising 2D photocatalysts. Consequently, our study not only unearths previously unexplored 2D potential photocatalysts but also introduces an effective screening methodology that may serve as a model for accelerating 2D materials discovery within a large chemical space.

INTRODUCTION

In future energy systems, hydrogen (H_2) is expected to have a prominent position as an energy carrier. As such, it is imperative that we produce hydrogen in an efficient and environmentally sustainable manner. In recent years, solar water splitting as a means of producing H_2 has gained significant interest; however, the efficacy of the process is heavily constrained by the performance of the photocatalysts.[1,2] Due to the maximal specific surface area, abundant active sites, short carrier migration distance, and appreciable sunlight absorption, 2D materials are regarded as promising candidates for serving as photocatalysts in the conversion of solar energy into valuable H_2 fuel.[3,4] Motivated by these advantages, the pursuit of new 2D photocatalysts has garnered increasing attention in recent years.[5–7] There is still a pressing need to find new 2D photocatalysts, given the stringent criteria they should meet, such as possessing a moderate band gap to maximize solar energy absorption and appropriate band edge positions to align with the water redox potentials.[8,9]

The field of material exploration has evolved into a new era characterized by big data and powerful ML algorithms, in relation to the remarkable growth in computing power and available computational data.[10,11] In recent years, several computational material databases for 2D materials have been developed,[12,13] which allow the use of ML techniques to uncover

[1]DIFFER – Dutch Institute for Fundamental Energy Research, De Zaale 20, Eindhoven 5612 AJ, the Netherlands.

[2]Materials Simulation and Modeling, Department of Applied Physics, Eindhoven University of Technology, Eindhoven 5600 MB, the Netherlands.

[3]Computational Chemical Physics, Faculty of Science and Technology and MESA+ Institute for Nanotechnology, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands.

*Corresponding author, email: s.er@differ.nl.

‡These authors contributed equally to this work.

intricate correlations within the material space and accurately predict material properties.[14,15] With the rapid prediction power of ML models, large chemical space of 2D materials can be screened based on the predictions of key properties and more accurate computational methods such as density functional theory (DFT) can be employed to expeditiously verify these predictions and create new training data for the models to extend the screening boundaries.[16–20]

In this work, we employ a combination of ML models and physics-based computational methods to uncover potential 2D semiconductor photocatalysts for solar water splitting from the V2DB database,[21] which comprises 316,505 likely stable 2D materials and their associated ML predicted properties. Figure 1 illustrates the three principal stages of our workflow, which are color-coded in blue, orange, and yellow. Each of these stages comprises multiple steps, which we describe in detail in the Results section. During the initial stage, we employed a combination of filtering criteria based on ML-predicted electronic properties and manual screening procedures to down-select promising photocatalysts. Next, we conduct PBE-level DFT computations, which encompass structural optimizations, band gap calculations, phonon dispersion calculations, and *ab initio* molecular dynamics (AIMD) simulations to identify stable 2D semiconductor photocatalysts. Finally, we utilize high-level calculation methods, including HSE06 and G_0W_0, to obtain accurate electronic properties and assess the photocatalytic water splitting potential of the top 2D semiconductor candidates. As a result of our screening approach, we identify 11 2D material candidates as promising photocatalysts for water-splitting. Our study not only puts forward new 2D semiconductor photocatalyst candidates for deeper investigation but also presents an effective ML-assisted physics-based computational screening approach that could inspire future searches for functional 2D materials.

RESULTS and DISCUSSION

ML-driven screening

All 2D materials in V2DB[21] were generated by enumeration of 22 different 2D crystal prototypes and 52 chemical elements from the periodic table. Through a series of ML model training based on a large amount of DFT (PBE level) data, V2DB contains a total of 316,505 potentially stable 2D materials and ML-predicted associated properties, including electronic band gap and band edge (PBE level) data. Compared to standard DFT methods, the GW method including many-body effects can give a more accurate description of electronic properties, band gaps and band edge positions in particular, are key properties for assessing the applicability of materials for solar water splitting. By applying a regression study between PBE and G_0W_0 electronic properties of 2D materials,[21] the associated predicted G_0W_0 band gaps and band edge positions were also collected in V2DB, which are used in the current work to screen potential semiconductor photocatalysts. The light-colored parts of violin maps, shown in Figure 2 (a) and Supplementary Information Figure S1, illustrate the distribution of V2DB materials ML-predicted G_0W_0 band gap values for different prototypes. For overall photocatalytic water splitting, the band gap of photocatalysts should exceed the water splitting free energy of 1.23 eV, and at the same time be smaller than 3 eV to efficiently harvest sunlight. Therefore, $1.23 \leq E_g^{ML} \leq 3$ eV is set as the first selection criterion in the current study, which is shown in Figure 2 (a) and Figure S1 as dashed red lines. In addition to band gap, photocatalysts for solar water splitting should also have appropriate band edges to straddle the water redox potentials. Specifically, the valence band maximum (VBM) should be lower

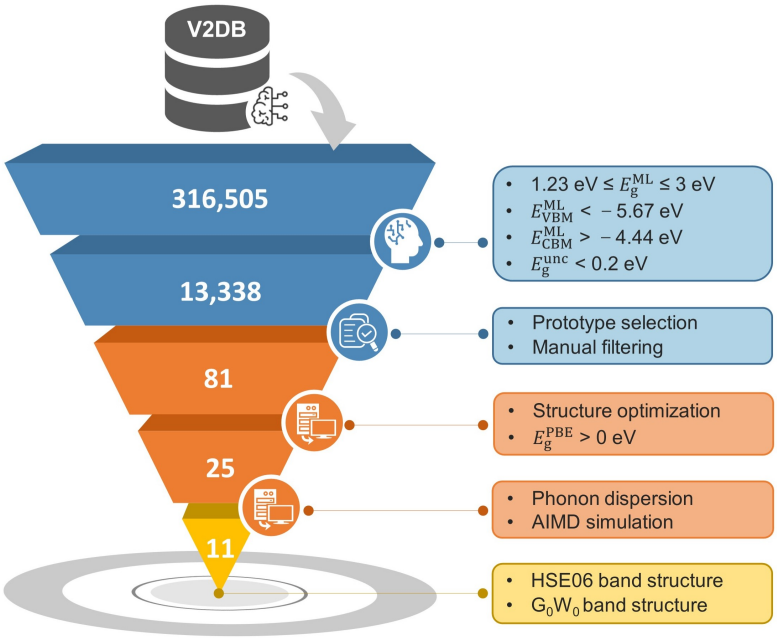

Figure 1. The workflow for scrutinizing V2DB as a possible source of 2D semiconductor photocatalysts for solar water splitting. The blue part contains the first screening layer (based on ML-predicted properties, from *Candidates-316,505* to *Candidates-13,338*) and the second screening layer (further manual selection, to *Candidates-81*). The orange part refers to computational screening, including the third layer (based on DFT-PBE level calculations, to *Candidates-25*) and the fourth layer (dynamical stability check, to *Candidates-11*). The yellow part represents the high-level (HSE06 and G_0W_0) electronic band structure calculations of the best 11 2D semiconductor candidates for solar water splitting.

than the oxidation potential of O_2/H_2O (−5.67 eV *vs* vacuum at pH=0), while the conduction band minimum (CBM) should be higher than the reduction potential of H^+/H_2 (−4.44 eV *vs* vacuum at pH=0), and these constitute our second and third criteria, respectively. The distribution of ML-predicted G_0W_0 band edges of all V2DB materials is shown in Figure 2 (b) with grey dots, as well as in Figures S2 and S3 split up according to the different 2D material prototypes. The black lines in these figures show the applied cutoffs for the band edges, while the diagonal lines denote the cutoffs for the band gap screening. Clearly, 2D materials located inside the triangle marked by these lines are ideal photocatalysts, which meet both the band gap and the band edge requirements.

The materials that constitute V2DB, however, are not uniformly distributed over its chemical space. Therefore, the properties of different types of materials are predicted with varying level of uncertainties. Although the accuracy in general of each ML model is specified in V2DB, uncertainties for the property prediction of each individual material are not provided. Therefore, in the present work we develop a meta-model that predicts the uncertainty of the previously predicted band gap values. The meta-model is based on a consensus of three different algorithms, namely Artificial Neural Network (ANN), Random Forest (RF), and eXtreme Gradient Boosting (XGB). To label our dataset and train the models, we collect the ML-predicted PBE band gap values from V2DB and the calculated PBE band gap values from C2DB[13], a computational 2D materials database, which is the source of ML training data for V2DB. Next, we quantify the uncertainty of the prediction as the absolute difference of these two values, which is calculated using Equation (1) below.

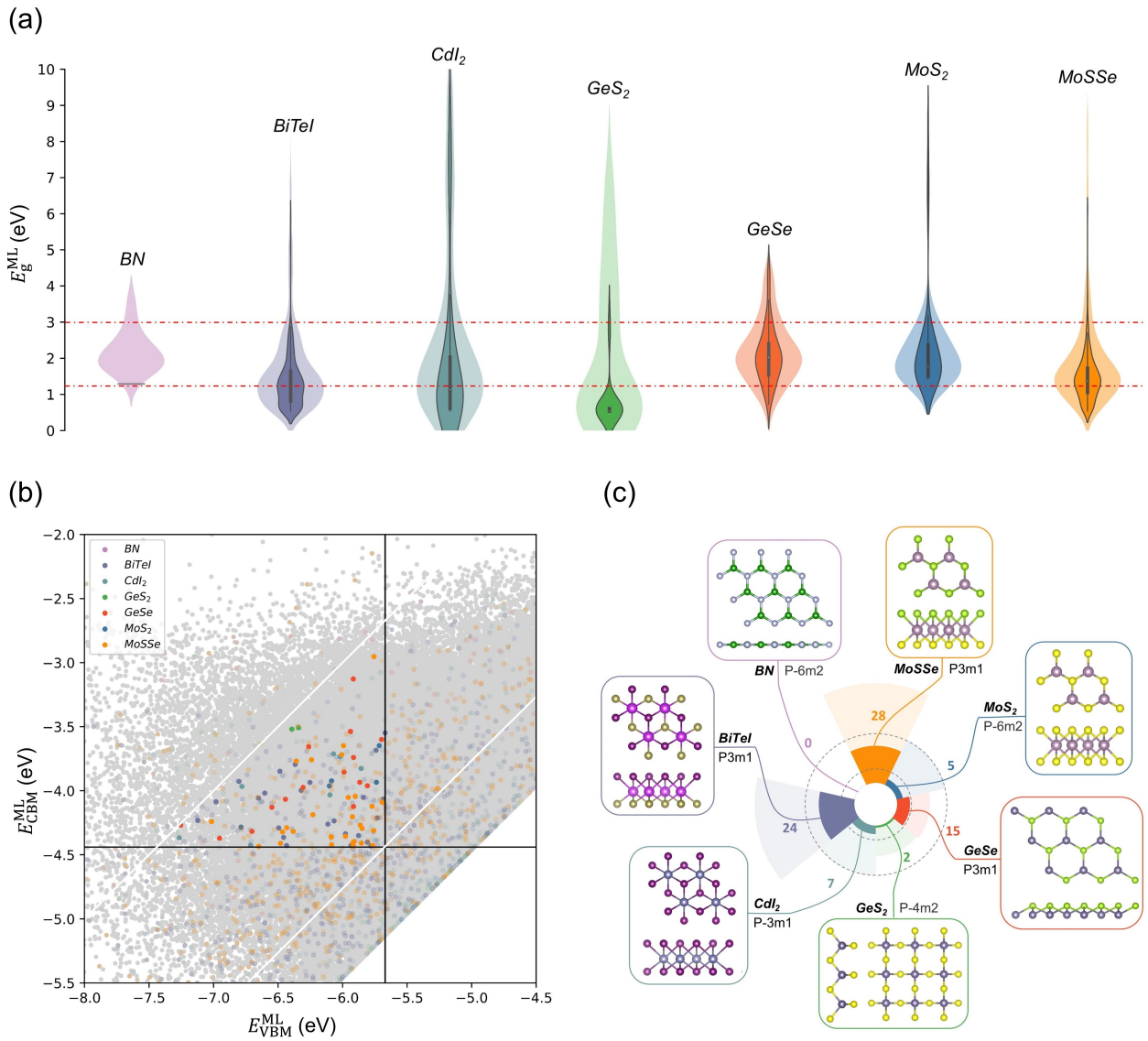

Figure 2. Screening based on ML-predicted properties. (a) The violin maps showing the distributions of ML-predicted band gap values for 2D materials from the seven prototypes identified in V2DB. For each prototype, the light-color shaded parts refer to all available candidates in V2DB, while the dark-color shaded parts refer to candidates with small values of ML-predicted band gap uncertainty ($E_g^{unc} < 0.2$ eV). The red-dashed lines denote the ML-predicted band gap selection criterion, $1.23 \leq E_g^{ML} \leq 3$ eV, applied in the current study. (b) The distribution of ML-predicted band edges in V2DB. The horizontal and vertical axes represent the VBM and CBM values, both relative to the vacuum level, respectively. The horizontal black line at -4.44 eV and the vertical black line at -5.67 eV represent the reduction potential of H^+/H_2 and the oxidation potential of O_2/H_2O at pH=0, respectively. Additionally, two diagonal white lines indicate band gaps of 1.23 and 3 eV. The grey dots represent all 2D materials from V2DB (*Candidates-316,505*). The light-colored dots represent candidates from seven selected prototypes, whereas the dark-colored dots in the triangle represent structures after the first-stage screening (*Candidates-81*). (c) The rose map and structural view of the seven prototypes. The light-colored regions of the rose map depict all structures belonging to the selected prototypes in V2DB, whereas the dark-colored areas and numbers correspond to *Candidates-81*.

$$E_g^{unc} = |E_g^{ML} - E_g^{DFT}| \quad (1)$$

where E_g^{ML} is the ML-predicted PBE band gap value, E_g^{DFT} is the calculated PBE band gap value, and E_g^{unc} is the uncertainty value of the ML-predicted band gap. Further technical information about the ML models, which are used to predict the band gap uncertainty data, are given in the Methods section. Using the trained meta-model, we predicate the band gap uncertainty for all 2D materials in V2DB. To determine the materials with accurately predicted band gap values, $E_g^{unc} < 0.2$ eV is used as the fourth filter criterion here. The darker colored parts, shown in Figure 2 (a) and Figure S1, denote the ML-predicted G_0W_0 band gap distribution of V2DB materials that satisfy the fourth criterion. After these four criteria are executed, the total number of 2D materials is reduced from 316,505 to 13,338 (*Candidates-13,338*).

In the second layer of the funnel, we impose two screening steps, including prototype selection and manual filtering, to effectively narrow down the candidate space. Taking into account the difficulties in experimental synthesis of the elementally complex 2D materials, here, the maximum number of atoms in a unit-cell is limited to three to screen relatively simple structures. Consequently, all prototypes with up to three atoms are preserved, which are *BN, BiTeI, CdI_2, GeS_2, GeSe, MoS_2*, and *MoSSe*. The structures of these selected prototypes are shown in Figure 2 (c). Additionally, the scatter plots for the ML-predicted G_0W_0 band edge values of the V2DB materials from these prototypes are shown in Figure 2 (b) using light colored dots. Considering that all atoms in the structures of these seven prototypes have usually more than one covalent bond with neighboring atoms, the existence of H atoms, with only one electron available for bonding, is not suited for formation of stable 2D-bonded material networks. Therefore, all structures that contain H atoms are also excluded in the current study. Then, manual filtering is performed to remove any remaining duplicate structures in the candidate list and the materials that are present in the ML training dataset. Additionally, 2D materials that have already been reported in literature are also excluded, as we aim here to find novel 2D photocatalysts. Table S1 shows examples of 2D structures that have been excluded during the manual filtering process. As a result of the filtering with the aforementioned criteria, the first stage screening ended up with 81 potential 2D photocatalysts for solar water splitting (*Candidates-81*), which successfully proceeded to the computational screening stage. The number of screened out candidates in *Candidates-81* from each prototype are shown in a rose map of Figure 2 (c). The ML-predicted G_0W_0 band edge distribution of these 81 candidates is illustrated in Figure 2 (b) in the form of dark colored dots. Besides the colored dots, there are a large number of grey dots located within the triangle shown in Figure 2 (b), which are as well candidate 2D materials for photocatalytic water splitting and awaiting to be explored in future studies. They all satisfy the band gap and band edge criteria applied in here but belong to the fifteen other prototypes in V2DB with more complex structure which are not considered in the current study.

Physics-based computational screening

The physics-based computational screening stage starts with structure relaxation of *Candidates-81*. Since in V2DB virtual 2D materials are generated without any geometrical optimization, it is essential to perform full structural optimization to find the most energy-favorable state of each candidate. The structure optimizations are carried out in multiple steps with increasing precision from low to high until the imposed convergence standards are reached for the total energy and forces exerting on each atom. The details of computational settings are provided in the Methods section. After the automated structural optimization, we

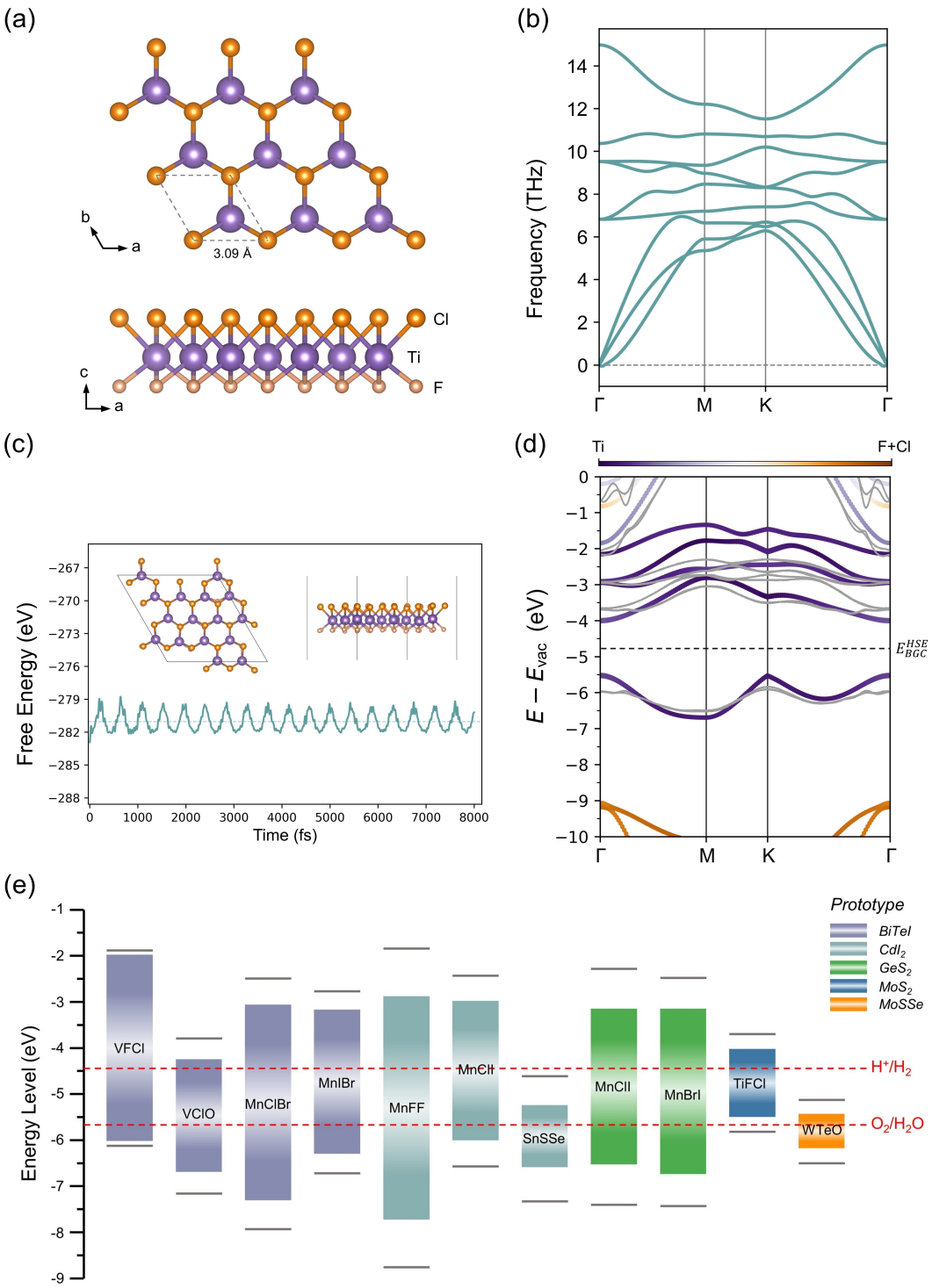

Figure 3. Computational screening and the final selection of 2D materials for water photocatalysis. (a) The top and side views of the TiFCl monolayer structure. (b) The phonon dispersion spectra of TiFCl. (c) The changes in the total free energy of TiFCl during an 8 ps AIMD simulation at 300 K, while the inset shows the top and side views of the equilibrium structure. (d) The element-projected fat band structure of TiFCl with respect to vacuum level, as calculated using the HSE06 functional. The color bar indicates the donation proportion of Ti, and F plus Cl elements. Grey lines depict G_0W_0 band structure relative to the vacuum level according to the HSE06 band gap center energy, E_{BGC}^{HSE} (indicated by the horizontal black dashed line). (e) The band alignments of *Candidates-11* relative to the vacuum level. The colored rectangular columns represent the HSE06 band edges, with different colors indicating the corresponding prototype. The G_0W_0 band edge positions are shown by the outer horizontal grey lines. The red dashed lines represent the reduction potential of H^+/H_2 and the oxidation potential of O_2/H_2O at pH=0.

manually check each structure to ensure they remain in reasonable 2D geometries. The majority of the candidates still maintain similar structures to their corresponding prototypes, although their lattice parameters and atomic positions undergo changes during structural optimization due to elemental differences in chemical compositions. For example, TiFCl monolayer preserves an *H*-phase sandwich configuration consistent with its prototype of MoS_2 as shown in Figure 3 (a). After determining the most energy-favorable geometry of *Candidates-81*, the electronic band gaps at PBE level, both with and without spin-orbit-coupling (SOC) effects, are calculated to identify the semiconductors. Due to the well-known underestimation of GGA-PBE functional, here, all candidates with a positive band gap value are kept. As a result, 25 semiconductors (*Candidates-25*) are identified after the third screening layer (Figure 1). The corresponding PBE-based electronic density of states (DOS), as calculated with the spin-polarization and SOC, are shown in Figures S4 and S5, respectively. In addition, the numerical values for the calculated band gaps are included in Table S2.

It is necessary to assess the dynamical stability of the candidate 2D structures before any time-consuming and labor-intensive experimental attempts. Hence, we calculate the phonon dispersion curves in the full Brillouin Zone (BZ) of *Candidates-25*, where a dynamically stable structure should have no imaginary frequencies. As an example, Figure 3 (b) shows the phonon dispersions of the TiFCl monolayer, where all dispersion curves of the six optical and three acoustic modes have positive frequencies only, demonstrating that TiFCl monolayer in its optimized structure is dynamically stable. However, one needs to stress that in first-principles phonon calculations of 2D materials, the flexural phonon branch, pertaining to out-of-plane motions, is sensitive to the computational parameters, and the existence of small imaginary values near the Γ point is a common issue.[22,23] The inaccuracy in properly describing such long wave length out-of plane motions is not relevant for the quantities discussed in this paper, however. Furthermore, 2D structures with small imaginary frequencies near Γ point can be stabilized by extrinsic effects, such as their interactions with substrates.[24] Therefore, here we only exclude structures that have large imaginary frequencies in the middle of the BZ, which high likely point to the structure being unstable.[25] The phonon calculation results of the *Candidates-25* are collected in Figure S6, in which the phonon stable 2D structures are labeled with red frames. Subsequently, we further estimate the thermodynamic stability of these phonon stable structures by AIMD simulations at T = 300 K. For example, Figure 3 (c) shows the total free energy during AIMD simulation of the TiFCl monolayer, indicating that the average energy remains at a fixed value. Moreover, there are no bond breaking or structural distortions in the final configuration. These results imply that the TiFCl monolayer should be stable in experiments at room temperature. Figure S7 includes a summary of all AIMD simulations results, revealing that two structures exhibiting noticeable distortions with the structure transforming completely from *H* to *T*-phase. Based on the phonon and AIMD results, 11 dynamically stable 2D semiconductors are screened out as photocatalyst candidates (*Candidates-11*). The corresponding structure configuration and lattice information of *Candidates-11* are presented individually in Figures S8-S18 (a, b).

In order to further investigate the electronic band structures of the selected potential 2D photocatalysts, the hybrid density function HSE06 with SOC correction is employed. The calculated HSE06 band gap (E_g^{HSE}) results of the *Candidates-11* materials are shown in Table 1. To visualize the contributions from each chemical element to the electronic states, we also

Table 1. The V2DB-ID, material formula, prototype information, band character, total magnetic moment (M_{total}; μ_B per formula unit), spin-polarized PBE band gap ($E_g^{PBE-spin}$), PBE band gap (E_g^{PBE}), HSE06 band gap (E_g^{HSE}), G_0W_0 band gap (E_g^{GW}), and G_0W_0 band edges with respect to the vacuum level of the best 2D material candidates, *Candidates-11*, for photocatalytic water splitting that have been identified in the current study. All the band gap and band edge values are shown in units of eV.

V2DB-ID	Material	Prototype	Band character	M_{total}	$E_g^{PBE-spin}$	E_g^{PBE}	E_g^{HSE}	E_g^{GW}	E_{VBM}^{GW}	E_{CBM}^{GW}
129	VFCl	*BiTeI*	Indirect	3.0	1.00	0.97	4.04	4.24	−6.12	−1.88
145	VClO	*BiTeI*	Direct	2.0	0.34	0.33	2.44	3.37	−7.16	−3.79
246	MnClBr	*BiTeI*	Indirect	5.0	1.67	1.62	4.25	5.44	−7.93	−2.49
286	MnIBr	*BiTeI*	Indirect	5.0	1.24	1.06	3.13	3.95	−6.72	−2.77
1375	MnFF	CdI_2	Direct	5.0	2.37	2.36	4.85	6.92	−8.76	−1.84
1392	MnClI	CdI_2	Indirect	5.0	1.18	0.97	3.03	4.15	−6.57	−2.43
1628	SnSSe	CdI_2	Indirect	0.0	0.99	0.84	1.45	2.71	−7.33	−4.61
1684	MnClI	GeS_2	Indirect	5.0	1.41	1.33	3.38	5.12	−7.40	−2.28
1686	MnBrI	GeS_2	Indirect	5.0	0.74	0.64	3.59	4.95	−7.43	−2.48
1781	TiFCl	MoS_2	Direct	0.0	1.15	1.12	1.48	2.12	−5.82	−3.70
3126	WTeO	*MoSSe*	Indirect	0.0	0.70	0.53	0.75	1.41	−6.52	−5.11

calculate the projected band structures of the *Candidates-11*, which are shown in Figures S8-S18 (c). One might expect that semiconductor where the VBM and CBM have different characters, *i.e.*, have contributions from different atom layers are more susceptible to spatial separation of the photo-excited carriers, which is an important step in initiating the photocatalytic reactions.[26,27] For instance, as shown in Figure S14 (c), the CBM of SnSSe has dominant contribution from the Sn atoms, while S and Se atoms contribute to most of the VBM. This indicates that the transition of electrons from VBM to CBM is accompanied by a spatial movement of the charges from the anions to the cations. Moreover, the character of band gap might affect different aspects of the photocatalytic efficiency as well. Direct band gap materials usually show a good solar energy absorption, while indirect band gap materials can have good performance in charge-carrier separation.[28,29]

The many-body G_0W_0 approach[30] with SOC correction based on the PBE wave-function is applied to calculate the G_0W_0 band gap (E_g^{GW}) of *Candidates-11*, and the corresponding results are listed in Table 1. Compared to the HSE06 band gap values, G_0W_0 generally increase the band gaps of the candidate materials. It should be noted, however, that the G_0W_0 band gaps calculated here are for free-standing 2D layers. These calculated values are likely to form an upper bound, as in an experimental situation the 2D layers will be adsorbed on the substrate, whose dielectric screening will reduce the band gap. In all cases PBE was used to generate the starting point for the G_0W_0 calculations, with the exception of VClO. In that particular case the PBE functional gives a qualitatively wrong description of the character of the bands around the band gap. This error is corrected by the PBE+U functional, see Figure S19, and we use this functional to generate the starting point for the G_0W_0 calculation on VClO. The G_0W_0 band gaps of the complete set of *Candidates-11* materials exceed the water splitting free energy of 1.23 eV. In addition, the calculated values for the three 2D semiconductors, including SnSSe, TiFCl and WTeO, are smaller than 3 eV, which suggests that these compounds can efficiently utilize the visible light spectrum. Admittedly, some candidates, such as MnFF and MnClBr, have relatively large band gap values at G_0W_0 level, which indicates that they are not perfectly

suitable for photocatalytic reactions. However, it may still be possible that these large band gap materials can exhibit promising photocatalytic performance under different conditions, including but not limited through the application of mechanical strain and electric field,[3] embedding in heterostructures.[31] It may also be possible to optimize such a material for one of the two half reactions of water splitting (either oxygen evolution, or hydrogen evolution), for instance by adjusting the pH values of the solution appropriately.[28]

To estimate the absolute values for the G_0W_0 band edges, we employ here the band gap center (BGC) model,[3,32] which has proven to be useful in previous studies.[26,33,34] As an example, Figure 3 (d) illustrates the G_0W_0 band structures of the TiFCl monolayer relative to the vacuum level obtained based on the HSE06 BGC energy. Table 1 and Figure 3 (e) give an overview of the calculated G_0W_0 band edge energy values for the *Candidates-11*. In Figure 3 (e), one can assess the water splitting performance of each candidate by comparing the VBM and CBM positions (grey lines) with the redox potentials of oxygen (O_2/H_2O) and hydrogen evolution (H^+/H_2) reactions at pH=0 (red dashed lines), respectively. Accordingly, the majority of *Candidates-11* have G_0W_0 band edges that straddle the redox potentials of water, which in principle make them promising candidates for overall photocatalytic water splitting. However, although the G_0W_0 method expands the band edges on the basis of HSE06 BGC energy, SnSSe and WTeO still only meet the VBM criteria, which indicates that they will be suitable only for the oxygen evolution reaction. Interestingly, the $E_{\mathrm{VBM}}^{\mathrm{GW}}$ of TiFCl is positioned under the oxygen evolution potential, which means it also has inherent capability for overall solar water splitting.

CONCLUSION

In summary, we employ an ML-aided physics-based computational funnel approach for the discovery of potential 2D solar water splitting photocatalysts from the V2DB database. The initial screening stage utilized ML-predicted electronic properties, including band gap and band edge positions, as well as a newly developed ML model to estimate the uncertainties in ML-predicted band gap values. We further narrowed down the list of candidates by selecting structures from prototypes with no more than three atoms per unit-cell, and excluding materials that have already been reported in the literature. The second stage of the funnel involves DFT calculations, beginning with structural optimization and band gap calculations at the PBE level to identify suitable semiconductors for photocatalytic water splitting. The stability of the remaining candidates is then checked by phonon dispersion and AIMD calculations. Following this two-stage screening process, we identify 11 dynamically stable 2D semiconductors as potential photocatalysts for solar water splitting. We further study these candidates by performing high-level hybrid-DFT (HSE06) and G_0W_0 calculations. As a result, all 11 candidates are found to have G_0W_0 gap values that exceed the water splitting free energy of 1.23 eV, while three 2D semiconductors SnSSe, TiFCl and WTeO have G_0W_0 gaps that are smaller than 3 eV, which implies an effectively use of visible sunlight for water photocatalysis. Furthermore, based on the HSE06 BGC energy, the G_0W_0 band edge positions are estimated for the best candidates. Nine out of 11 2D materials are found to have band edges that straddle around the water redox potentials, whereas the remaining two materials are found to be suitable only for the oxygen evolution reaction. Therefore, several newly identified 2D photocatalysts in this work hold promise for future validation. Thus, this research could potentially inspire further exploration of novel 2D materials by

effectively utilizing both physics-based calculations and ML models.

METHODS

Machine Learning Algorithms

We employed a feature vector and three different ML algorithms with the optimized parameters as explained below. Our consensus ML model consists of a combination of these individual ML models, with the final predictions being obtained by calculating the arithmetic average of the predictions made by each of these models.

Feature Vector Configuration

To represent the materials in the latent space we constructed a feature vector by amalgamating three different vectors using the same methodology as described in the V2DB study[21]. The first one is the prototype vector, which is a one-hot vector containing the prototype information of the materials. The second vector is the chemical composition vector, which comprises details about the chemical elements present in the materials, along with the composition of these elements. Lastly, the electronegativity vector contains float data values of electronegativity, calculated using the geometric mean of the electronegativity of positively and negatively charged atoms in the compound, and has a length of 2.

Artificial Neural Network (ANN) Model Configuration

We used scikit-learn[35] to train the ANN models and performed parameter optimization using grid search with the configurations given below. To optimize the parameters, we employed 10-fold cross-validation. The parameters selected for the final ANN model that demonstrated the best performance are shown below in bold, while the remaining parameters were set to their default values.

- Activation function: (**relu**, tanh)
- Alpha: (0.025, 0.05, **0.1**, 0.2)
- Max iteration: (**100**, 200, 400)
- Hidden Layers: [(100),(200),(400),(800),(100,10),(200,10),**(400,10)**,(800,10),(100,20),(200,20),(400,20),(800,20)]

Random Forest (RF) Model Configuration

We used scikit-learn[35] library to train the RF models. We used a total of 1000 estimators for building the model. The remaining parameters were set to default values.

Extreme Gradient Boosting (XGB) Model Configuration

We used xgboost[36] library to train the XGB models. Using grid search, we optimized the parameters of the XGB models with the configurations given below. To optimize the given parameters, we used 10-fold cross-validation. The selected parameters for the final XGB model that exhibited the best performance are indicated in bold below, whereas the default values were used for the remaining parameters.

- Max depth: (4, 6, 8, **10**)
- Learning rate: (0.2, **0.3**, 0.4)

- Min child weight: (**1**, 2)
- gamma: (**0**, 0.1, 0.2, 0.3, 0.4)
- Colsample bytree: (0.3, 0.5, **0.7**)

DFT calculations

All DFT calculations were performed with the Vienna *ab initio* Simulation Package (VASP),[37,38] and the results were handled by the VASPKIT package.[39] We employed the frozen-core projector augmented wave (PAW) method and Perdew-Burke-Ernzerhof (PBE) functional within the generalized gradient approximation (GGA).[40,41] To avoid interactions with the mirror image, a sufficiently large vacuum space of 15 Å was added perpendicular to the surface in the *c* direction for each 2D structure. The electronic wavefunction was expanded using a plane-wave basis with a kinetic energy cutoff of 500 eV. All the geometric structures were visualized using the VESTA package.[42]

In the structure optimization calculations, the convergence criteria were set to be 10^{-5} eV in energy between the consecutive relaxation steps and 0.01 eV/Å in force remaining on atoms. The 2D Brillouin zones were sampled using a $12\times12\times1$ Γ-centered *k*-point mesh. For the density of state (DOS) calculations, the effects of magnetism and spin-orbital coupling (SOC) corrections were taken into account, and denser *k*-point grids of $21\times21\times1$ were used.

The phonon dispersion curves were calculated using density functional perturbation theory (DFPT) as implemented in the PHONOPY code.[43] Supercells of $5\times5\times1$ unit cells were employed for this purpose. The AIMD simulations were performed employing the canonical ensemble (*NVT*) at 300 K for 8 ps, at a time step of 1 fs, and using supercells of $4\times4\times1$. The temperature was controlled using the Nosé-Hoover method. Additionally, for magnetic materials, the phonon and AIMD calculations involved the spin-polarization effects.

HSE06 calculations

To achieve accurate band structure calculations, the Heyd–Scuseria–Ernzerhof (HSE06) hybrid density functional was employed, which incorporates 25% exact Hartree−Fock (HF) exchange.[44] In these calculations, the effects of SOC were included, while the band edge positions relative to the vacuum level were accurately computed using the electrostatic potential alignments.

G_0W_0 calculations

The one-shot G_0W_0 calculations were performed, for which the quasi-particle energies were obtained using DFT-PBE wave functions.[30] For these calculations, $12\times12\times1$ *k*-grids were applied, which is consistent with several recent high-throughput studies.[19,45,46] The energy convergence and energy cutoff for the response function were set to 10^{-8} between consecutive electronic steps and 100 eV, respectively. To ensure accurate results, 160 empty bands were included and the correction effect of SOC was taken into account in each computation. The maximally localized Wannier functions were fitted to the quasi-particle band structures using the WANNIER90 package.[47]

According to the BGC model,[3,32] the band edges at the G_0W_0 level are calculated as follows

$$E_{\mathrm{VBM}}^{\mathrm{GW}} = E_{\mathrm{BGC}}^{\mathrm{HSE}} - \frac{1}{2}E_{\mathrm{g}}^{\mathrm{GW}} \tag{2}$$

$$E_{\mathrm{CBM}}^{\mathrm{GW}} = E_{\mathrm{BGC}}^{\mathrm{HSE}} + \frac{1}{2}E_{\mathrm{g}}^{\mathrm{GW}} \tag{3}$$

where $E_{\mathrm{BGC}}^{\mathrm{HSE}}$ is the band gap center energy calculated using the HSE06 hybrid functional, and $E_{\mathrm{g}}^{\mathrm{GW}}$ is the band gap calculated using the G_0W_0 approximation.

DATA AVAILABILITY

The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information files.

ACKNOWLEDGEMENTS

We thank Dr. Engin Torun for fruitful discussion. We acknowledge funding from the initiative "Computational Sciences for Energy Research" of Shell and NWO (Grant No 15CSTT05). Y. W. acknowledges financial support from China Scholarships Council (Grant No 202006930008). This work was sponsored by NWO Exact and Natural Sciences for the use of supercomputer facilities.

AUTHOR CONTRIBUTIONS

Y. W. conducted all the physics-based computation. M. C. S. conducted ML models of band gap uncertainty. S. E. devised and supervised the project. All authors discussed the results, reviewed and edited the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ASSOCIATED CONTENT

The Supplementary Information is available for this paper. The ML-predicted properties and physics-based calculation results are collected in SI. Violin maps of ML-predicted G_0W_0 band gaps (Fig S1); scatter plot of ML-predicted G_0W_0 band edges distribution of V2DB 2D structures (Fig S2, S3); examples of 2D structures excluded in the manual filtering step (Table S1). The calculated PBE DOS results with spin-polarization (Fig S4) and SOC effects (Fig S5), and their summary (Table S2). The phonon (Fig S6) and AIMD (Fig S7) calculation results. The results of final selection of 11 candidates are individually shown in Fig S7-S18, including structural views, calculated HSE06 and G_0W_0 band structures. The band structures of VClO as calculated with (a) PBE; (b) PBE+U, (c) G_0W_0; (d) G_0W_0+U methods (Fig S19).

References

1. Chen, X., Shen, S., Guo, L. & Mao, S. S. Semiconductor-based photocatalytic hydrogen generation. *Chem. Rev.* **110**, 6503–6570 (2010).

2. Qu, Y. & Duan, X. Progress, challenge and perspective of heterogeneous photocatalysts. *Chem. Soc. Rev.* **42**, 2568–2580 (2013).

3. Singh, A. K., Mathew, K., Zhuang, H. L. & Hennig, R. G. Computational screening of 2D materials for photocatalysis. *J. Phys. Chem. Lett.* **6**, 1087–1098 (2015).

4. Ganguly, P. *et al.* 2D nanomaterials for photocatalytic hydrogen production. *ACS Energy Lett.* **4**, 1687–1709 (2019).

5. Wang, X. *et al.* A metal-free polymeric photocatalyst for hydrogen production from water under visible light. *Nat. Mater.* **8**, 76–80 (2009).

6. Xiang, Q., Yu, J. & Jaroniec, M. Synergetic effect of MoS_2 and graphene as cocatalysts for enhanced photocatalytic H_2 production activity of TiO_2 nanoparticles. *J. Am. Chem. Soc.* **134**, 6575–6578 (2012).

7. Su, T., Shao, Q., Qin, Z., Guo, Z. & Wu, Z. Role of interfaces in two-dimensional photocatalyst for water splitting. *ACS Catal.* **8**, 2253–2276 (2018).

8. Li, Y., Li, Y.-L., Sa, B. & Ahuja, R. Review of two-dimensional materials for photocatalytic water splitting from a theoretical perspective. *Catal. Sci. & Technol.* **7**, 545–559 (2017).

9. Fu, C.-F., Wu, X. & Yang, J. Material design for photocatalytic water splitting from a theoretical perspective. *Adv. Mater.* **30**, 1802106 (2018).

10. Agrawal, A. & Choudhary, A. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Mater.* **4**, 053208 (2016).

11. Chen, A., Zhang, X. & Zhou, Z. Machine learning: Accelerating materials development for energy storage and conversion. *InfoMat* **2**, 553–576 (2020).

12. Zhou, J. *et al.* 2DMatPedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches. *Sci. Data* **6**, 1–10 (2019).

13. Haastrup, S. *et al.* The Computational 2D Materials Database: high-throughput modeling and discovery of atomically thin crystals. *2D Mater.* **5**, 042002 (2018).

14. Moosavi, S. M., Jablonka, K. M. & Smit, B. The role of machine learning in the understanding and design of materials. *J. Am. Chem. Soc.* **142**, 20273–20287 (2020).

15. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).

16. Tran, K. & Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. *Nat. Catal.* **1**, 696–703 (2018).

17. Masood, H., Toe, C. Y., Teoh, W. Y., Sethu, V. & Amal, R. Machine learning for accelerated discovery of solar photocatalysts. *ACS Catal.* **9**, 11774–11787 (2019).

18. Jin, H. *et al.* Discovery of novel two-dimensional photovoltaic materials accelerated by machine learning. *J. Phys. Chem. Lett.* **11**, 3075–3081 (2020).

19. Kumar, R. & Singh, A. K. Chemical hardness-driven interpretable machine learning approach for rapid search of photocatalysts. *npj Comput. Mater.* **7**, 1–13 (2021).

20. Yin, H. *et al.* The data-intensive scientific revolution occurring where two-dimensional materials meet machine learning. *Cell Reports Phys. Sci.* **2**, 100482 (2021).

21. Sorkun, M. C., Astruc, S., Koelman, J. V. A. & Er, S. An artificial intelligence-aided virtual screening recipe for two-dimensional materials discovery. *NPJ Comput. Mater.* **6**, 1–10 (2020).

22. Cahangirov, S., Topsakal, M., Aktürk, E., Şahin, H. & Ciraci, S. Two- and one-dimensional honeycomb structures of silicon and germanium. *Phys. Rev. Lett.* **102**, 236804 (2009).

23. Wang, J., Yip, S., Phillpot, S. R. & Wolf, D. Crystal instabilities at finite strain. *Phys. Rev. Lett.* **71**, 4182–4185 (1993).

24. Zólyomi, V., Drummond, N. & Fal'Ko, V. Electrons and phonons in single layers of hexagonal indium chalcogenides from ab initio calculations. *Phys. Rev. B* **89**, 205416 (2014).

25. Mounet, N. *et al.* Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nat. Nanotechnol.* **13**, 246–252 (2018).

26. Zhao, P., Liang, Y., Ma, Y., Huang, B. & Dai, Y. Janus chromium dichalcogenide monolayers with low carrier recombination for photocatalytic overall water-splitting under infrared light. *J. Phys. Chem. C* **123**, 4186–4192 (2019).

27. Ju, L., Bie, M., Tang, X., Shang, J. & Kou, L. Janus WSSe monolayer: an excellent photocatalyst for overall water splitting. *ACS Appl. Mater. & Interfaces* **12**, 29335–29343 (2020).

28. Jin, H. *et al.* Data-driven systematic search of promising photocatalysts for water splitting under visible light. *J. Phys. Chem. Lett.* **10**, 5211–5218 (2019).

29. Peng, R., Ma, Y., Huang, B. & Dai, Y. Two-dimensional janus PtSSe for photocatalytic water splitting under the visible or infrared light. *J. Mater. Chem. A* **7**, 603–610 (2019).

30. van Schilfgaarde, M., Kotani, T. & Faleev, S. Quasiparticle self-consistent *GW* theory. *Phys. Rev. Lett.* **96**, 226402 (2006).

31. Zhang, X. *et al.* Computational screening of 2D materials and rational design of heterojunctions for water splitting photocatalysts. *Small Methods* **2**, 1700359 (2018).

32. Toroker, M. C. *et al.* First principles scheme to evaluate band edge positions in potential transition metal oxide photocatalysts and photoelectrodes. *Phys. Chem. Chem. Phys.* **13**, 16644–16654 (2011).

33. Liang, Y., Huang, S., Soklaski, R. & Yang, L. Quasiparticle band-edge energy and band offsets of monolayer of molybdenum and tungsten chalcogenides. *Appl. Phys. Lett.* **103**, 042106 (2013).

34. Zhuang, H. L. & Hennig, R. G. Computational search for single-layer transition-metal dichalcogenide photocatalysts. *J. Phys. Chem. C* **117**, 20440–20445 (2013).

35. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

36. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).

37. Becke, A. D. & Edgecombe, K. E. A simple measure of electron localization in atomic and molecular systems. *J. Chem. Phys.* **92**, 5397–5403 (1990).

38. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).

39. Wang, V., Xu, N., Liu, J.-C., Tang, G. & Geng, W.-T. VASPKIT: A user-friendly interface facilitating high-throughput computing and analysis using VASP code. *Comput. Phys. Commun.* 108033 (2021).

40. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953 (1994).

41. Perdew, J. P. *et al.* Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Phys. Rev. B* **46**, 6671–6687 (1992).

42. Momma, K. & Izumi, F. VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. *J. Appl. Crystallogr.* **44**, 1272–1276 (2011).

43. Togo, A. & Tanaka, I. First principles phonon calculations in materials science. *Scripta Materialia* **108**, 1–5 (2015).

44. Perdew, J. P. *et al.* Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys. Rev. Lett.* **100**, 136406 (2008).

45. Zhao, P. *et al.* Two-dimensional III2-VI3 materials: promising photocatalysts for overall water splitting under infrared light spectrum. *Nano Energy* **51**, 533–538 (2018).

46. Yang, X., Singh, D., Xu, Z., Wang, Z. & Ahuja, R. An emerging janus mosete material for potential applications in optoelectronic devices. *J. Mater. Chem. C* **7**, 12312–12320 (2019).

47. Pizzi, G. *et al.* Wannier90 as a community code: new features and applications. *J. Physics: Condens. Matter* **32**, 165902 (2020).