npj Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Unsupervised discovery of thin-film photovoltaic materials from unlabeled data

Zhilong Wang[1,2], Junfei Cai[1,2], Qingxun Wang[1,2], SiCheng Wu[1,2] and Jinjin Li[1]✉

Quaternary chalcogenide semiconductors (I_2-II-IV-X_4) are key materials for thin-film photovoltaics (PVs) to alleviate the energy crisis. Scaling up of PVs requires the discovery of I_2-II-IV-X_4 with good photoelectric properties; however, the structure search space is significantly large to explore exhaustively. The scarcity of available data impedes even many machine learning (ML) methods. Here, we employ the unsupervised learning (UL) method to discover I_2-II-IV-X_4 that alleviates the challenge of data scarcity. We screen all the I_2-II-IV-X_4 from the periodic table as the initial data and finally select eight candidates through UL. As predicted by ab initio calculations, they exhibit good optical conversion efficiency, strong optical responses, and good thermal stabilities at room temperatures. This typical case demonstrates the potential of UL in material discovery, which overcomes the limitation of data scarcity, and shortens the computational screening cycle of I_2-II-IV-X_4 by ~12.1 years, providing a research avenue for rapid material discovery.

npj Computational Materials (2021)7:128 ; https://doi.org/10.1038/s41524-021-00596-4

INTRODUCTION

Solar energy is the most important basic energy among all types of renewable energy[1,2]. The technologies that convert solar energy to electrical power (such as photovoltaic (PV) generation and photoelectrochemical generation, will receive more attention in multi-functional clean energy sources[3–5]. Practical thin-film PV cells are based on quaternary chalcogenides (I_2-II-IV-X_4) of sphalerite crystals such as CdTe and Cu(In, Ga)(S, Se)$_2$ (CIGSSe), which is cheaper to process and exhibit competitive performance levels compared to conventional crystal silicon-based PVs. Their battery power conversion efficiency (PCE) exceeds 20% at present[6,7]. However, these materials require expensive or rare elements (In, Te), or even toxic (Cd), severely limiting their large-scale development. Kesterite $Cu_2ZnSn(S, Se)_4$ (CZTSSe), is a potential thin-film PV material, in which the In and Ga in CIGSSe are replaced with Zn and Sn, and its record PCE of 12.6% is significantly lower than that of CdTe/CIGSSe[8]. One possible reason for this is the antisite disorder in the kesterite structure, which significantly affects the open-circuit voltage and device performance. CZTSSe, where the smaller Zn of is replaced by Ba with a larger ionic radius ($Cu_2BaSnS_{4-x}Se_x$(CBTSSe)) to ease the antisite disorder, has demonstrated better performance in PV in comparison with CZTSSe[9–12]. Continuing this process, replacing Cu with Ag in CBTSSe to form $Ag_2BaSnSe_4$ (ABTSe), and replacing Ba with Sr to form $Cu_2SrSnSe_4$ (CSTSSe), yields materials that have recently shown great promise with respect to thin-film PV applications[13,14].

These cases enlighten us that it is worth exploring the wider space of I_2-II-IV-X_4 chalcogenide semiconductors (where I-, II-, and IV-sites are occupied by the different oxidation states of the cations and X-site is a chalcogenide anion)[15–17], to identify earth-abundant, environmentally friendly thin-film PV materials inspired by existing compounds. A possible path toward the discovery of I_2-II-IV-X_4 materials is to synthesize or theoretically calculate the properties of massive sets of potential structures in terms of element substitution (e.g., I = Li^+, Cu^+, or Ag^+; II = Ba^{2+} or Sr^{2+}; IV = Sn^{4+} or Ge^{4+}; X = O^{2-}, S^{2-}, or Se^{2-}), and then screen for materials with good electro-optical properties. For example, semiconductors with band gaps (E_g) (particularly with direct band gaps) in the visible wavelength region (0.9–1.6 eV, the range of optimum optical conversion efficiency), and strong optical responses in the visible spectrum, are considered promising thin-film PVs[18,19]. Unfortunately, such an approach is impractical because of the high costs and long cycles of the necessary experimentation and is also not amenable to high-throughput computing due to excessive computational costs. Finding a way to quickly discover I_2-II-IV-X_4 chalcogenide semiconductors is an important challenge for current research, which is of great significance for identifying thin-film PVs and further improving PCE.

With the rise of machine learning (ML) applications, data-driven approaches to material design and selection have promoted the development of materials science. ML methods extend far beyond the limitations of other current electronic structure analysis methods, to investigate novel, emergent phenomena originating from the complexity of the physical systems[20–23]. ML technologies, such as deep neural networks (DNNs)[24,25], support vector machines (SVMs)[26,27], and random forest (RF)[28,29] algorithms, have made remarkable achievements in materials science. Ding et al. used over 10^4 catalytic samples to design non-noble metal electrocatalytic proton exchange membrane fuel cells[30]. Based on the ML model, Ali et al. achieved fast recovery of the cubic structure in mixed cation perovskite thin films from high-throughput calculation database[31]. Moreover, there are some remarkable reports on the application of ML methods in PV materials[23,32–34]. However, the applications of ML in these systems use supervised learning, the biggest imperfection of which is that it still requires an adequate data set to ensure the accuracy of predictions. According to our current knowledge base, the data set of I_2-II-IV-X_4 chalcogenides is still quite scarce, and existing supervised learning models are unable to predict properties based on structures, and a relevant ML model has not been reported

[1]National Key Laboratory of Science and Technology on Micro/Nano Fabrication, Shanghai Jiao Tong University, Shanghai, China. [2]Department of Micro/Nano Electronics, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. ✉email: lijinjin@sjtu.edu.cn

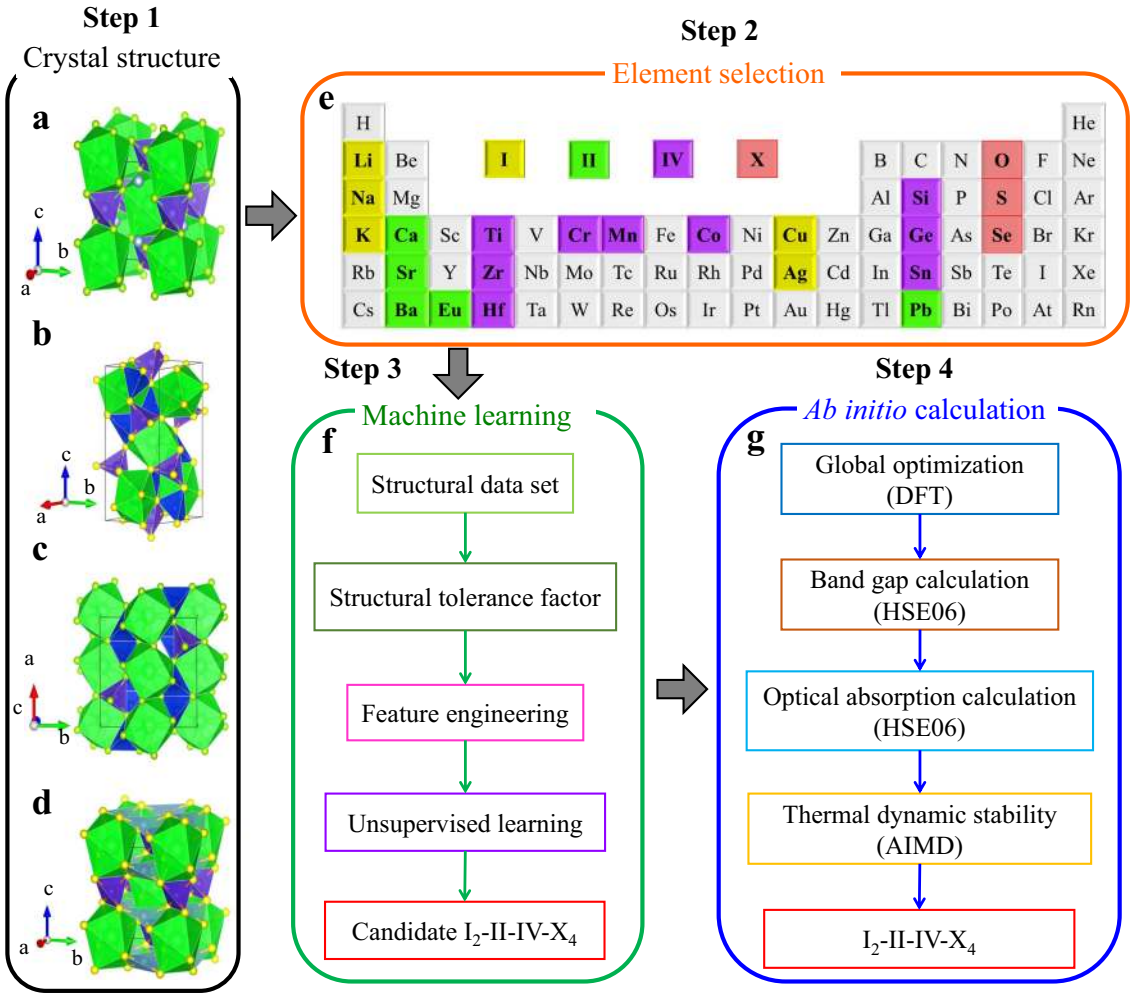

Fig. 1 Schematic of the unsupervised discovery of I_2-II-IV-X_4 chalcogenides. **a** Crystallographic view of $Ag_2BaSnSe_4$ ($I222$). **b** Crystallographic view of Cu_2BaSnS_4 ($P3_1$). **c** Crystallographic view of $Cu_2BaSnSe_4$ ($Ama2$). **d** Crystallographic view of Ag_2BaGeS_4 ($I\bar{4}2m$). **e** Element selection of I (yellow), II (green), IV (purple), and X (light red) from the periodic table. **f** Workflow of an unsupervised guided discovery of candidate I_2-II-IV-X_4 chalcogenides. **g** Further accurate verification of the I_2-II-IV-X_4 chalcogenides through ab initio calculations. The gray arrows represent the sequential workflow.

with such a small data set. Consequently, a more sensible strategy to overcome the limited data available is urgently needed.

In this study, we proposed an unsupervised learning (UL) model with unlabeled data and apply it to a representative case of exploring the I_2-II-IV-X_4 chalcogenides for thin-film PV materials. Based on the structure of I_2-II-IV-X_4 chalcogenides and recombination of elements from the periodic table, a total of 2700 structures (containing 27 identified materials) with four different space groups were selected as the initial data set, and 1520 candidates were screened out based on the tolerance factor. We used an agglomerative hierarchical clustering (AHC) algorithm[35] to accomplish UL, and proposed a descriptor representing the sums and differences of elemental properties to cluster I_2-II-IV-X_4 chalcogenides. Our unsupervised model clusters I_2-II-IV-X_4 chalcogenides into one group with suitable E_g, while the other groups of materials had larger E_g values. Based on the high-precision Heyd–Scuseria–Ernzerhof calculations (HSE06) method, we quantitatively calculated the E_g and optical absorption coefficient[36,37] of the selected compounds, and successfully discovered eight I_2-II-IV-X_4 chalcogenides with good electro-optical properties (Ag_2BaTiS_4, $Ag_2BaTiSe_4$, Ag_2BaCrS_4, $Ag_2BaSiSe_4$, Ag_2BaZrS_4, $Ag_2BaZrSe_4$, $Ag_2BaHfSe_4$, and $Cu_2BaMnSe_4$). We further demonstrated that these chalcogenides have good thermal stabilities at room temperature using ab initio molecular dynamic (AIMD) simulations. The proposed AHC-UL model bypasses the challenge of data scarcity in traditional ML methods and effectively avoids extremely long computational and experimental cycles. Based on the recombination of all elements in the periodic table, the proposed model reduces the screening period of I_2-II-IV-X_4 chalcogenides by ~12.1 years. We hope that the eight I_2-II-IV-X_4 chalcogenides proposed from 2700 unknown compounds will be served as promising thin-film PVs to significantly improve PCE.

RESULTS

Workflow of material discovery

The workflow for unsupervised discovery of I_2-II-IV-X_4 chalcogenides for thin-film PVs is illustrated in Fig. 1, including four modules: determination of crystal structures (Fig. 1a–d)[38], element selection from the periodic table (Fig. 1e), the establishment of the ML model (Fig. 1f), and ab initio calculation (Fig. 1g). In this procedure, considering that the I_2-II-IV-X_4 chalcogenide has four different space groups, $I222$, $P3_1$, $Ama2$, and $I\bar{4}2m$, we selected one structure from each space group as initial structures, as shown in Fig. 1a (see elemental sites and structures in Supplementary Fig. 1). Then, the proper site elements were selected according to the oxidation state and coordination number from the periodic table, where I = Li^+, Na^+, K^+, Cu^+, or Ag^+; II = Ca^{2+}, Sr^{2+}, Ba^{2+}, Eu^{2+}, or

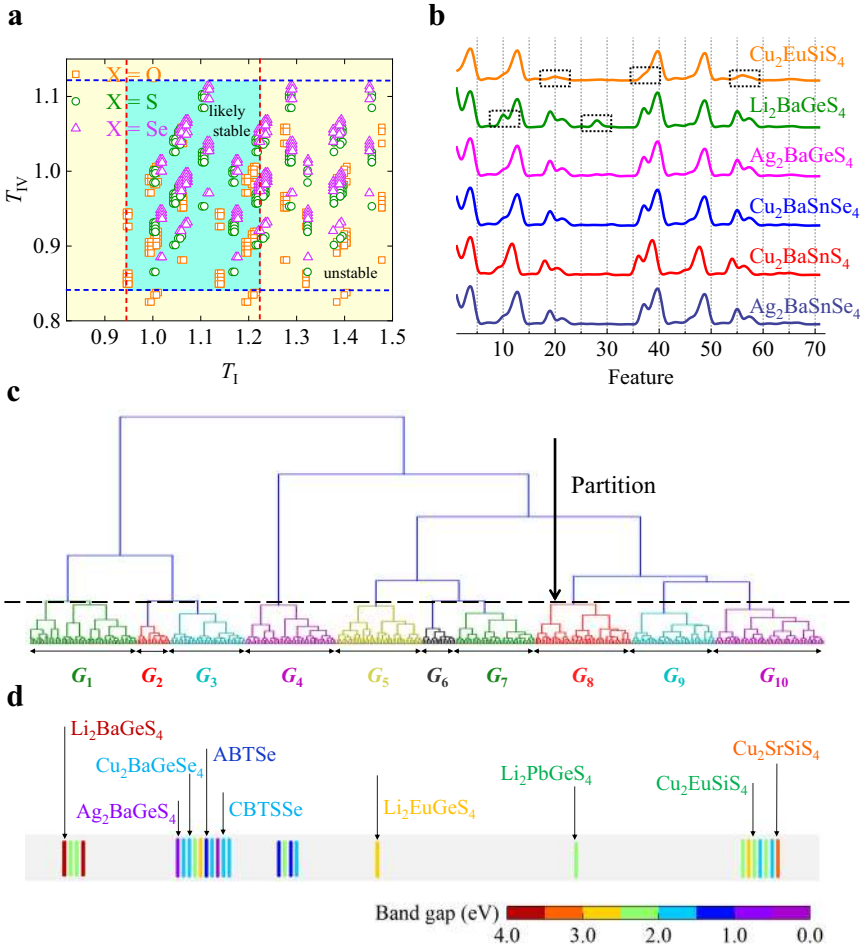

Fig. 2 Unsupervised learning of I_2-II-IV-X_4 semiconductors. a Plot of tolerance factors (T_I and T_{IV}) of 675 I_2-II-IV-X_4 compounds, where 380 likely stable compounds are in the cyan area. **b** Computed features of six selected compounds based on SDEPs, dotted boxes show significant differences in features. **c** Bottom-up tree diagram (dendrogram) generated by the agglomerative hierarchical clustering (AHC) method. The dashed line shows the position where all compounds are partitioned into ten groups, marked as G_1–G_{10} from left to right and distinguished by different colors. **d** Mapping the dendrogram to the band gaps, revealing the grouping of 27 known I_2-II-IV-X_4 semiconductors (Supplementary Table 1). The color bar shows the scale of band gap.

Pb^{2+}; IV = Ti^{4+}, Zr^{4+}, Hf^{4+}, Cr^{4+}, Mn^{4+}, Co^{4+}, Si^{4+}, Ge^{4+}, or Sn^{4+}; X = O^{2-}, S^{2-}, or Se^{2-}, as shown in Fig. 1e. Thus, 2700 different structures (675 compounds without considering four space groups) were generated as the initial data set for the ML module. We used the tolerance factor (T_f), to make preliminary judgments regarding the structural stabilities[16], leaving 1520 I_2-II-IV-X_4 chalcogenides to be further studied (380 compounds without considering space groups). Next, strong relationships between compounds were established by feature engineering, and the 380 structures were clustered based on the AHC-UL algorithm. Finally, 26 candidates covering two space groups ($I222$ and $P3_1$) were selected from one of the ten groups, corresponding to Fig. 1f. As shown in Fig. 1g, ab initio calculations were performed to predict the electro-optical properties and evaluate the thermal stabilities of the 26 candidates, and eight I_2-II-IV-X_4 chalcogenides were identified as promising thin-film PV materials. The step-by-step screening process is discussed in Supplementary Fig. 2 and Supplementary Note 1.

Unsupervised learning of I_2-II-IV-X_4 chalcogenides

The process used here for UL is shown in Fig. 1f, consisting of three parts: data set, feature engineering, and algorithm[39,40]. First, there are five, five, nine, and three elements to choose for the I-, II-, IV-, and X-sites, respectively, which can form 675 different compounds, and 2700 different structures with four space groups. Then, T_f (T_I and T_{IV}) were applied to make a preliminary judgment about the structural stabilities of these compounds. As shown in Fig. 2a, the T_f plot of 675 compounds distinguished by X-site is presented, where 380 compounds with $0.94 < T_I < 1.22$ and $0.84 < T_{IV} < 1.11$ (cyan area in Fig. 2a, see Supplementary Fig. 3) are potentially stable (see more details of T_f in Methods). Thus, a data set of 380 I_2-II-IV-X_4 compounds without considering space groups were selected for the next UL clustering.

Feature engineering, which can transform raw random data into model training data to be closely related to the output attributes, and determines the upper limit of the ML model. The goal of this work is to determine I_2-II-IV-X_4 chalcogenides with good electro-optical properties. The band gap E_g is a basic parameter of electronic properties, which needs to be considered first. Therefore, we need to build a feature set to create a strong relationship between the compounds and electronic properties. The factors affecting the E_g are complex, but to train the model, the feature set must be limited. From previous works[21,41–43], we found that the elemental properties of materials have good mapping relationships with their band gaps. Therefore, nine elemental properties were selected, including the atomic number (Z), group number (g), covalent radius (R_{cov}), and first ionization energy (E_{ie}). The list of all elemental properties is provided in Supplementary Table 2. To construct the feature vectors, we proposed a

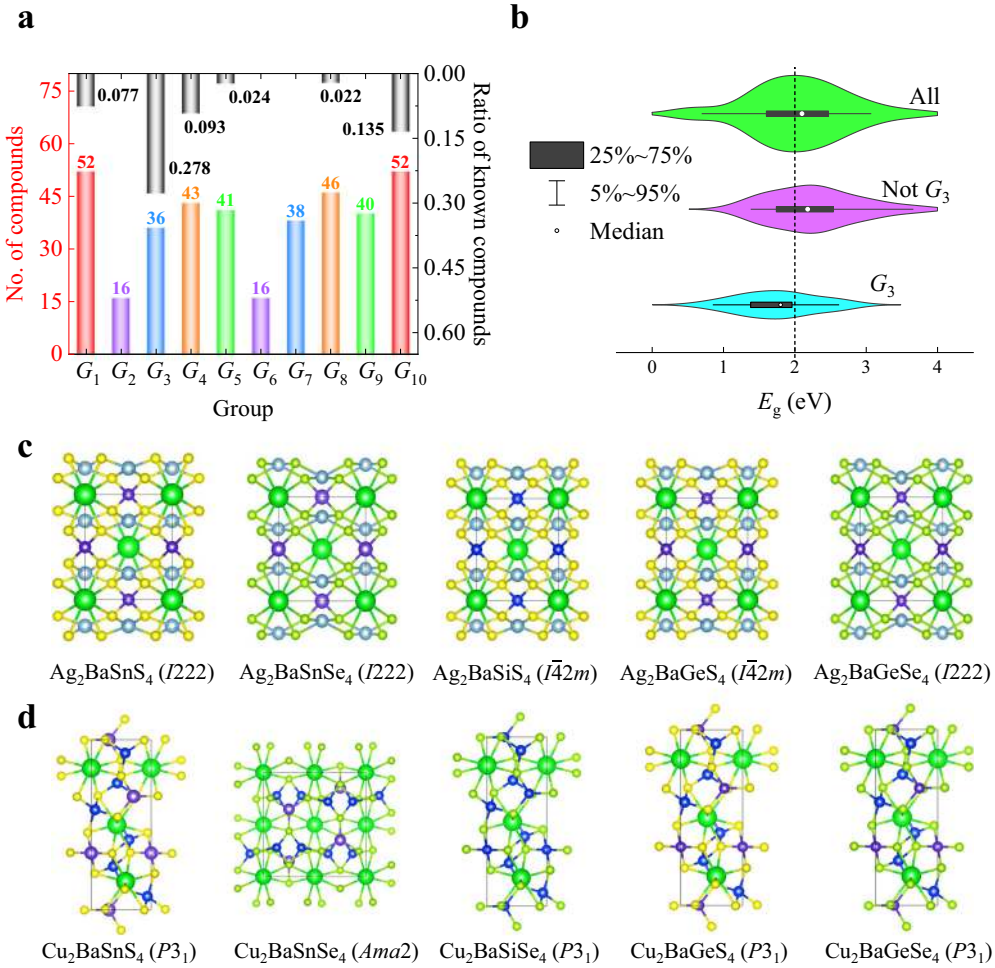

Fig. 3 Further analysis of group partitioning by unsupervised learning. a A statistical graph of the number of compounds in each group (left), and the ratio of known compounds in each group (right). **b** Violin plots of E_g of 27 known compounds, ten known compounds in G_3, and other 17 known compounds not clustered in G_3. The outer shells of the violins bound all data, narrow horizontal lines bound 90% of the data, thick horizontal lines bound 50% of the data, and white dots represent medians. The dashed line shows the position of $E_g = 2.0$ eV. **c** Five known I_2-II-IV-X_4 semiconductors in G_3, where I = Ag^+. **d** Five known I_2-II-IV-X_4 semiconductors in G_3, where I = Cu^+.

descriptor using the sums and differences of elemental properties (SDEPs for short) for I_2-II-IV-X_4 compounds, and constructed a 72-dimensional SDEP in this way for each I_2-II-IV-X_4 compound (see Methods and Supplementary Fig. 4). The feature plots of the six selected compounds (Cu_2EuSiS_4, Li_2BaGeS_4, Ag_2BaGeS_4, $Cu_2BaSnSe_4$, Cu_2BaSnS_4, and $Ag_2BaSnSe_4$) are shown in Fig. 2b. As the input of clustering, these feature curves look very similar and play an important role in clustering results. From Fig. 2b, the feature curves of the selected six structures are very similar, but there are still significant differences (see dotted boxes), such as the tenth and 20th features, which are also the key to clustering.

Based on SDEPs, the AHC-UL algorithm[35,44] was used to cluster the 380 I_2-II-IV-X_4 compounds. The bottom-up tree diagram (dendrogram) generated by the AHC is presented in Fig. 2c, where an appropriate partition line is selected and the 380 compounds are classified into ten groups (from G_1, G_2, …, to G_{10}). The different colors in Fig. 2c correspond to different groups. More details about the position of the partition line are discussed in Supplementary Note 2 and Supplementary Fig. 5. The grouping shows a good quality of clustering as different groups are well differentiated, and the SDEPs share similar characteristics within the same groups (Supplementary Fig. 4). Details of the 72-dimensional feature vector for each compound are provided in Supplementary Table 3. Therefore, a visible clustering of I_2-II-IV-X_4 compounds can be found using AHC. From Fig. 2c, d, most of the known I_2-II-IV-X_4 compounds with $E_g < 2.0$ eV are clustered into G_3 in the dendrogram (Fig. 2d), including ten structures of $Ag_2BaGeSe_4$ ($E_g = 0.85$ eV), ABTSe ($E_g = 1.42$ eV), $Cu_2BaGeSe_4$ ($E_g = 1.88$ eV), and CBTSSe ($E_g = 1.96$ eV), etc. In addition, G_1, including $Li_2BaGeSe_4$ ($E_g = 2.4$ eV) and Li_2BaSnS_4 ($E_g = 3.07$ eV), are all known compounds with $E_g > 2.0$ eV. For G_5 and G_8, each contains only one known compound with $E_g > 2.0$ eV, $Li_2EuGeSe_4$ ($E_g = 2.54$ eV), and Li_2PbGeS_4 ($E_g = 2.41$ eV), respectively. Thus, as confirmed by the above correlations between the groups and band gaps, our proposed SDEPs and UL model can capture the chemical and physical relations of the electronic properties of I_2-II-IV-X_4 chalcogenides. Besides, we also performed the K-means method to cluster I_2-II-IV-X_4 compounds, the comparison of AHC and K-means are provided in Supplementary Fig. 6, Supplementary Table 4, and Supplementary Note 3.

Physical insights from unsupervised learning

The clustering of I_2-II-IV-X_4 chalcogenides by SDEPs provides physical insights into the understanding of compounds exhibiting useful stabilities, proper electronic properties, and suitable crystal structures. We counted the number of compounds in each group, including both unknown compounds and 27 known compounds (Fig. 3a). G_1 and G_{10} had the most compounds at 52 each, while G_2 and G_6 contained the fewest compounds at 16 each, indicating

that a targeted study of these groups would significantly narrow down the initial scope, which contained 380 compounds with four space groups. In addition, we generalize and summarize the ratios of known compounds in these groups (Fig. 3a and Supplementary Table 4). Remarkably, G_3 contains 36 compounds (Supplementary Table 4), of which ten are known compounds, accounting for 27.8%, which is much higher than the ratios of known compounds in other groups. In terms of stability, the structures in G_3 are likely to be more stable, as the 27 known compounds have been experimentally synthesized. This is an obvious implication that the unknown 26 compounds in G_3 deserve further targeted study.

As shown in Fig. 3b, the violin plot of the ten known compounds in G_3 showed a significantly lower E_g, with an average of 1.75 eV and a median of 1.80 eV, and the majority of the 17 known compounds outside of G_3 have significantly higher E_g, with an average of 2.27 eV and a median of 2.18 eV. This shows great potential for the discovery of I_2-II-IV-X_4 compounds with lower E_g among the 26 unknown compounds in G_3 with good electro-optical properties. Moreover, from Supplementary Table 5 and Supplementary Fig. 7, we found that the 36 compounds in G_3 showed the same I-site of Ag^+ or Cu^+, II-site of Ba^{2+}, and X-site of S^{2-} or Se^{2-}, revealing the dependence of their stabilities and electronic properties on their elemental properties. T hese discoveries from UL have important guiding significance for the design of I_2-II-IV-X_4 compounds with stable and good electronic properties.

In addition to the discovery of elemental and electronic properties, we also found patterns in the crystal structures of the four space groups. The crystal structures of five known compounds with I = Ag^+ and II = Ba^{2+} are shown in Fig. 3c. These structures contain two kinds of similar space groups: $I\bar{4}2m$ (40%) and $I222$ (60%), where the I-X_4 tetrahedra are flattened and share edges with the II-X_8 dodecahedra, do not bear any resemblance to the square antiprisms observed in the $P3_1$ and $Ama2$ structures (Fig. 1a and Supplementary Fig. 1). In Fig. 3d, where the other five known compounds with I = Cu^+ and II = Ba^{2+} are presented, they happen to contain two other types of space groups, $Ama2$ (20%) and $P3_1$ (80%). These phenomena indicate that the ions at the I-site have an effect on the crystal structures, and the stable structures can be hopefully obtained when different chemical formulas are combined with suitable space groups (e.g., when I = Ag^+, the space group of $I222$ is better, while for I = Cu^+, the space group of $P3_1$ is better). To further study the 26 unknown compounds in G_3 and narrow down the screening, we focused on two main space groups; when Ag^+ is at the I-site, we selected $I222$, while when Cu^+ is at the I-site, we selected $P3_1$. As a result, the scope of exploration narrowed from 1520 structures to 26 structures (see the list in Supplementary Table 5).

From Supplementary Fig. 4, the dashed boxes show the significant different features in G_3, which are 10th, 19th, 56th features, indicating that the difference of dipole polarizability between I, II, and IV-site elements and the difference of atomic number between I, or II, or IV and X-site elements are important to electronic property of the I_2-II-IV-X_4 compounds (Supplementary Table 3). In addition to speeding up the screening process, UL can estimate important features for guiding the designing of high-performance I_2-II-IV-X_4 compounds. Therefore, the supervised learning method may be further developed to accurately analyze the importance of features.

Electro-optical properties of eight I_2-II-IV-X_4 chalcogenides

We performed ab initio calculations to predict the electro-optical properties of the 26 I_2-II-IV-X_4 structures in G_3. First, geometric structure optimizations were performed, and the crystal structures remained in good geometric arrangements. To achieve high precision prediction, we used the high-level HSE06 calculation, which is considered to be close to the experimental results[37,45–48], to calculate the electronic and optical properties of the screened 26 candidates. From Table 1 and Fig. 4, 20 structures were determined to be semiconductors ($E_g > 0$), with direct (^{D}) or indirect (^{I}) band gaps. The band structures and density of states

Table 1. Ab initio calculations for 20 semiconductors from G_3.

Compound	Space group	Lattice constant (Å)	Lattice angle (°)	Volume (Å^3)	E_g (eV)
Ag_2BaTiS_4	$I222$	$a = 6.50$, $b = 7.46$, $c = 8.28$	$\alpha = \beta = \gamma = 90$	395.75	^{D}1.42
$Ag_2BaTiSe_4$	$I222$	$a = 6.77$, $b = 7.74$, $c = 8.54$	$\alpha = \beta = \gamma = 90$	447.37	^{D}1.18
Ag_2BaCrS_4	$I222$	$a = 6.54$, $b = 7.18$, $c = 8.16$	$\alpha = \beta = \gamma = 90$	383.74	^{I}0.70
Ag_2BaZrS_4	$I222$	$a = 6.59$, $b = 7.49$, $c = 8.41$	$\alpha = \beta = \gamma = 90$	415.37	^{I}1.93
$Ag_2BaZrSe_4$	$I222$	$a = 6.83$, $b = 7.86$, $c = 8.70$	$\alpha = \beta = \gamma = 90$	466.91	^{I}1.60
Ag_2BaHfS_4	$I222$	$a = 6.58$, $b = 7.49$, $c = 8.40$	$\alpha = \beta = \gamma = 90$	414.09	^{D}2.13
$Ag_2BaHfSe_4$	$I222$	$a = 6.82$, $b = 7.89$, $c = 8.67$	$\alpha = \beta = \gamma = 90$	466.67	^{D}1.76
$Ag_2BaSiSe_4$	$I222$	$a = 7.02$, $b = 7.44$, $c = 8.36$	$\alpha = \beta = \gamma = 90$	436.22	^{D}1.33
Cu_2BaTiS_4	$P3_1$	$a = b = 6.29$, $c = 15.73$	$\alpha = \beta = 90$, $\gamma = 120$	538.00	^{I}2.27
$Cu_2BaTiSe_4$	$P3_1$	$a = b = 6.58$, $c = 16.48$	$\alpha = \beta = 90$, $\gamma = 120$	619.00	^{I}2.02
Cu_2BaCrS_4	$P3_1$	$a = b = 6.21$, $c = 15.32$	$\alpha = \beta = 90$, $\gamma = 120$	511.18	^{I}2.24
$Cu_2BaCrSe_4$	$P3_1$	$a = b = 6.49$, $c = 16.17$	$\alpha = \beta = 90$, $\gamma = 120$	590.19	^{I}2.30
Cu_2BaMnS_4	$P3_1$	$a = b = 6.18$, $c = 15.29$	$\alpha = \beta = 90$, $\gamma = 120$	505.29	^{I}2.09
$Cu_2BaMnSe_4$	$P3_1$	$a = b = 6.48$, $c = 15.99$	$\alpha = \beta = 90$, $\gamma = 120$	580.86	^{I}0.87
Cu_2BaZrS_4	$P3_1$	$a = b = 6.39$, $c = 16.00$	$\alpha = \beta = 90$, $\gamma = 120$	565.61	^{I}2.69
$Cu_2BaZrSe_4$	$P3_1$	$a = b = 6.67$, $c = 16.83$	$\alpha = \beta = 90$, $\gamma = 120$	648.93	^{I}2.40
Cu_2BaCoS_4	$P3_1$	$a = b = 6.21$, $c = 15.00$	$\alpha = \beta = 90$, $\gamma = 120$	500.60	^{I}2.54
Cu_2BaHfS_4	$P3_1$	$a = b = 6.36$, $c = 16.09$	$\alpha = \beta = 90$, $\gamma = 120$	563.16	^{I}3.01
$Cu_2BaHfSe_4$	$P3_1$	$a = b = 6.67$, $c = 16.82$	$\alpha = \beta = 90$, $\gamma = 120$	647.77	^{I}2.73
Cu_2BaSiS_4	$P3_1$	$a = b = 6.20$, $c = 15.46$	$\alpha = \beta = 90$, $\gamma = 120$	515.16	^{I}3.19

Optimized lattice constants, lattice angles, volumes, and band gaps are presented.
^{D} the direct band gap, ^{I} the indirect band gap.

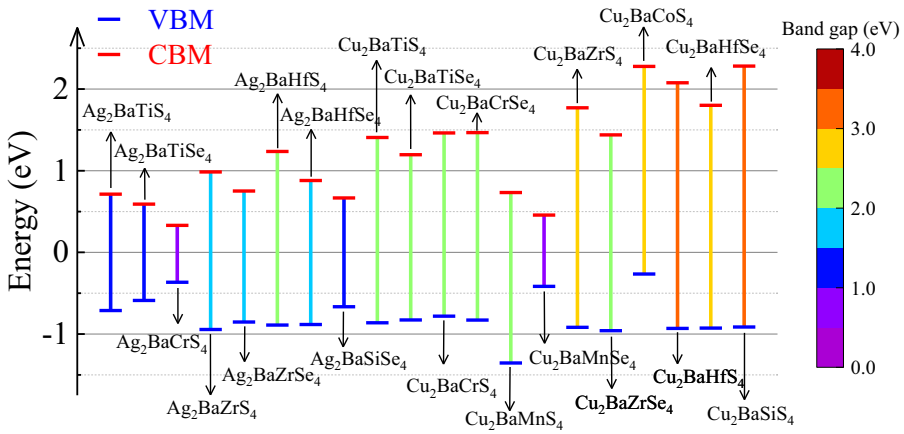

Fig. 4 Edges of VBMs and CBMs for 20 I_2-II-IV-X_4 semiconductors. The color bar (from purple to red) shows the scale of band gap.

(DOSs) of 20 structures are shown in Fig. 5 and Supplementary Figs. 8–23. From the predicted band structures, 15 structures have indirect band gaps, most of which come are the structures with I = Cu^+. The other five structures have direct band gaps, all from the structures with I = Ag^+ (1.42 eV for Ag_2BaTiS_4, 1.18 eV for $Ag_2BaTiSe_4$, 2.13 eV for Ag_2BaHfS_4, 1.76 eV for $Ag_2BaHfSe_4$, and 1.33 eV for $Ag_2BaSiSe_4$).

The edges of the conduction band minimums (CBMs) and valence band maximums (VBMs) of 20 semiconductors are shown in Fig. 4. We found that when I = Ag^+, CBMs and VBMs tend to be symmetric and have smaller E_g, which may lead to more information for applications such as photocatalysis and bipolar tubes, in addition to thin-film PVs. When I = Cu^+, most CBMs and VBMs are asymmetric, and CBMs are generally farther from Fermi level than VBMs. For IV-site, taking Ti, Zr, and Hf as an example, we found that with the increase of atomic number, the positions of the CBMs and VBMs deviate from the Fermi level, leading to the wider band gaps. The corresponding DOSs are also presented in Fig. 5 and Supplementary Figs. 8–23, where CBMs are dominated by I- and X-site atoms, while VBMs are mainly influenced by IV-site atoms. All the CBMs and VBMs show that II-site atoms (Ba^{2+}) do not dominate the scene, which is consistent with the fact that II-site atoms are indistinctive in the 26 unknown structures (which are all Ba^{2+}) from G_3. The above results indicate the importance of elements in regulating CBMs and VBMs to impact the E_g of I_2-II-IV-X_4 structures, and UL can effectively cluster similar elemental and conductivity characteristics into one group.

It is worth noting that there are four structures (1.42 eV for Ag_2BaTiS_4, 1.18 eV for $Ag_2BaTiSe_4$, 1.33 eV for $Ag_2BaSiSe_4$, and 1.60 eV for $Ag_2BaZrSe_4$, as shown in Fig. 4) having E_g between 0.9 and 1.6 eV, which is the range of optimal optical conversion efficiency[18]. In particular, Ag_2BaTiS_4, $Ag_2BaTiSe_4$, and $Ag_2BaSiSe_4$ have direct E_g, this means that the electron transitions do not require phonon release or absorption; As a result, electrons and holes are more likely to recombine. They may be used as potential thin-film PVs. Moreover, Ag_2BaCrS_4, Ag_2BaZrS_4, $Ag_2BaHfSe_4$, and $Cu_2BaMnSe_4$ have band gaps around 0.9 or 1.6 eV, are also likely to have high optical conversion efficiency, we also took them into account. Since PV suitability is the primary motivation for examining these properties, more analysis of the optical properties of these eight structures is necessary. The calculated absorption coefficients (α) of these eight structures based on the HSE06 functional are presented in Fig. 6 and Supplementary Fig. 24. They all show strong optical responses ($\alpha > 10^5\ cm^{-1}$) in the visible spectrum (1.65–3.26 eV, the colorful background in Fig. 6), and the absorption coefficients are largely isotropic in this range, showing only minor variations among ∥ *a*, ∥ *b*, and ∥ *c* directions, potentially indicating that there is only a small performance dependence on film orientation for thin-film PVs. The optimal band gaps and desired optical absorptions of the eight I_2-II-IV-X_4 chalcogenides show that they have great promise as thin-film PVs and exhibit good performance.

In addition to the eight outstanding candidates for thin-film PVs, the proposed method screens out 12 other I_2-II-IV-X_4 chalcogenide semiconductors that are far away from the range of 0.9–1.6 eV (the corresponding band structures and DOSs are provided in Supplementary Figs. 8–23), which are also likely to play important roles in PV devices, even in other fields (photocatalysis, sensors, detectors, etc.). There are many suggested improvements to properly adjust their band gaps or optical properties to achieve superior performance.

Our UL model not only overcomes the problem of data scarcity, but also greatly shortens the cycle of I_2-II-IV-X_4 chalcogenide discovery, positively differing high-throughput calculations by previous works[49,50]. In this work, 27 known structures were excluded from the initial 1520 I_2-II-IV-X_4 chalcogenides, and 26 candidates were finally screened out. In terms of computational screening, each structure required 260,464 s on a 24-CPU supercomputer (Supplementary Fig. 25) through the high-precision HSE06 method, meaning that the present work saves ~12.1-year computational cycles for 1467 structures. This will provide research export and method for the next generation of thin-film PVs. In addition, structural features also have an important impact on the band gap prediction[51], but we did not consider them in the clustering due to the lack of data. With the further development of high-throughput computing, more and more I_2-II-IV-X_4 materials will be available, at which point we will be able to establish supervised learning models to accurately predict band gaps using classification or regression methods.

Thermodynamic stabilities of eight I_2-II-IV-X_4 chalcogenides

For the eight I_2-II-IV-X_4 chalcogenides with good electro-optical properties, their thermodynamic stability should be evaluated in addition to the preliminary stability determined by UL and geometric structure optimization in order to facilitate their practical application. Therefore, we performed AIMD to evaluate the thermodynamic stabilities of the screened I_2-II-IV-X_4 candidates. As shown in Fig. 7 and Supplementary Fig. 26, the total energies of all the systems fluctuate within a very small range without a clear drop or rise during the simulations at 300 K. The crystal structure snapshots were extracted by an interval of 1.0 ps, without obvious expansion or contraction. Moreover, formation enthalpy is an important criterion to test crystal stability[34,52], thus we also calculated the formation enthalpies for eight I_2-II-IV-X_4

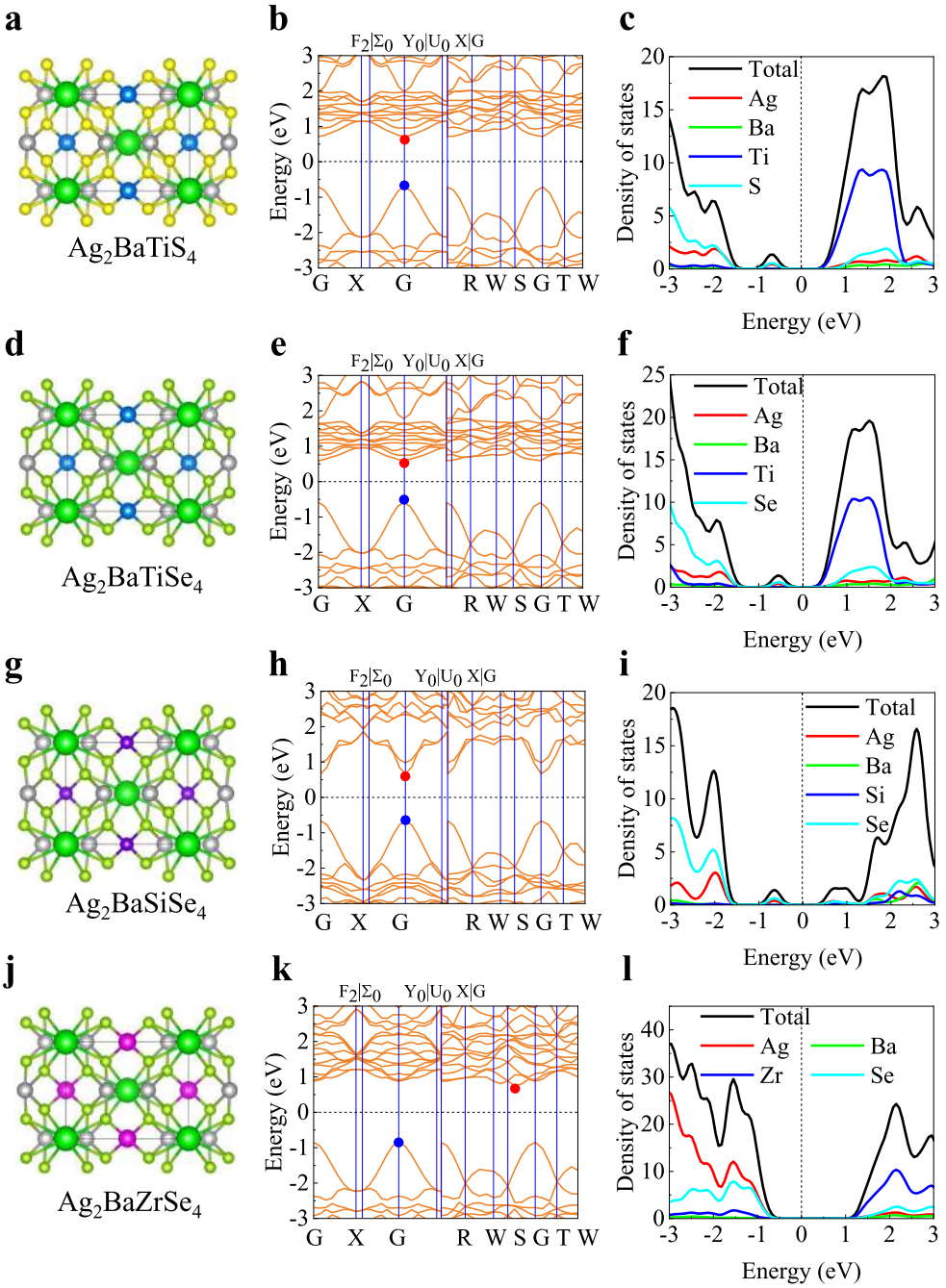

Fig. 5 Four I_2-II-IV-X_4 chalcogenides with E_g in the range of 0.9–1.6 eV, predicted by HSE06 calculations. The optimized crystal structures, band structures and density of states of Ag_2BaTiS_4, $Ag_2BaTiSe_4$, $Ag_2BaSiSe_4$, and $Ag_2BaZrSe_4$ are shown in **a–c** Ag_2BaTiS_4, **d–f** $Ag_2BaTiSe_4$, **g–i** $Ag_2BaSiSe_4$, **J–l** $Ag_2BaZrSe_4$. In the band structures, the red dots present CBMs, while the blue dots present VBMs. The electronic structures of the other four I_2-II-IV-X_4 chalcogenides are provided in Supplementary Figs. 13, 20, 21, and 23.

candidates with two host space groups of *I222* and $P3_1$. As shown in Supplementary Fig. 27, all structures with a space group of *I222* show the lower formation enthalpies than the structures with space group of $P3_1$, especially seven structures with $I = Ag^+$ show the lower formation enthalpies than experimental Ag_2BaSnS_4, indicating their good stability. It is noting that the selected $Cu_2BaMnSe_4$ with a space group of $P3_1$ has a slightly higher formation enthalpy than $Cu_2BaMnSe_4$ with a space group of *I222* and experimental Cu_2BaSnS_4, which may be metastable since the small difference in formation enthalpy and its smooth energy fluctuations during AIMD. The individual energy values (isolated atoms and bulk structures) are provided in Supplementary Table 6.

The above results indicate that the eight I_2-II-IV-X_4 chalcogenides can maintain the integrity of their crystal structures and good thermal stabilities at room temperature. This indicates that the eight structures selected in this work are stable. We expect that they can be further widely applied in thin-film PVs with good performance.

DISCUSSION

I_2-II-IV-X_4 chalcogenides have become important materials for thin-film PVs. The discovered I_2-II-IV-X_4 chalcogenides meet the criteria for earth-abundance and environmental friendliness, and

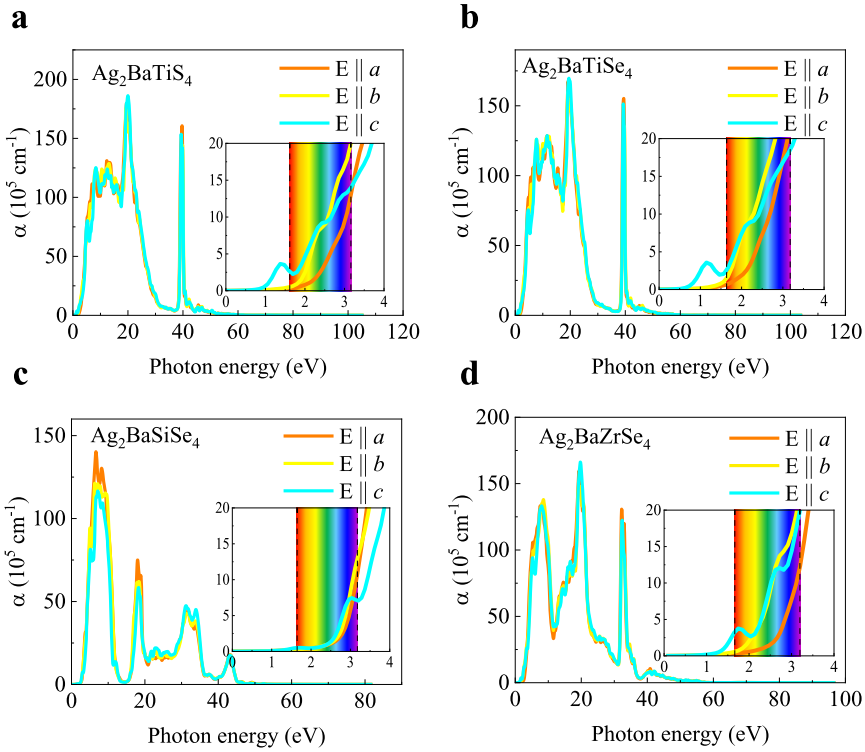

Fig. 6 Calculated optical absorption coefficients of four I_2-II-IV-X_4 chalcogenides. a Ag_2BaTiS_4, **b** $Ag_2BaTiSe_4$, **c** $Ag_2BaSiSe_4$, **d** $Ag_2BaZrSe_4$. E || *a* parallel to reciprocal *a* axis, E || *b* parallel to reciprocal *b* axis, and E || *c* parallel to the reciprocal *c* axis. The optical absorption coefficients of the other four I_2-II-IV-X_4 chalcogenides are provided in Supplementary Fig. 24.

demonstrate great potential for improving PV performance. However, traditional approaches, such as experimental synthesis and high-throughput computing, are limited due to the long-time cycle, and even the reliable ML model is impeded by the scarcity of material property data.

In summary, our achievements are as follows: (1) We propose an accessible descriptor of SDEPs based on the isolated elemental properties to obtain the feature vector of I_2-II-IV-X_4 compounds, which can be expanded to other material systems. (2) eight I_2-II-IV-X_4 compounds (Ag_2BaTiS_4, $Ag_2BaTiSe_4$, Ag_2BaCrS_4, $Ag_2BaSiSe_4$, Ag_2BaZrS_4, $Ag_2BaZrSe_4$, $Ag_2BaHfSe_4$, and $Cu_2BaMnSe_4$) with optimal band gaps, desired optical absorptions, and practical thermal stabilities at room temperatures were selected out of 2700 original structures by UL. They demonstrate great potential as thin-film PVs. (3) Because each structure requires an average of 260,464 s with a 24-CPU supercomputer, our method significantly reduced the scope for screening and calculation (from 2700 structures to 1520 structures, to 26 structures), dramatically shortening the computational cycle of material discovery by ~12.1 years. (4) This study demonstrates the potential of UL in material discovery, thus surmounting the obstacle of data scarcity, which may lead to important ideas and methods for the future discovery of materials.

Furthermore, we hope that the eight candidates revealed in this work will be synthesized experimentally for the preparation and application of thin-film PVs. For the other 16 I_2-II-IV-X_4 semiconductors identified, they are also of high research value in different fields according to their band gaps. In our subsequent work, we will focus on finding better descriptors to explore more precise quantitative laws of I_2-II-IV-X_4 structures, such as the expanded Shannon radii, cell volume[33]. Meanwhile, this work is a typical case of UL in material discovery, we look forward to its vigorous development in materials science.

METHODS

Tolerance factor

Tolerance factors serve as descriptors for phase stability with quaternary I_2-II-IV-X_4 semiconductors, which can be used for structure prediction in an empirically driven learning model[16]. For I_2-II-IV-X_4 materials, two dimensionless tolerance factors (T_I and T_{IV}) describing the geometric relations have been derived recently. The formula is as follows,

$$T_I = \sqrt{\frac{4+\sqrt{2}}{3}}\frac{r_I + r_X}{r_{II} + r_X} \quad (1)$$

$$T_{IV} = \sqrt{\frac{4+\sqrt{2}}{3}}\frac{r_{IV} + r_X}{r_{II} + r_X} \quad (2)$$

where r_I, r_{II}, r_{IV}, and r_X are the ionic radii of the I, II, IV, and X elements, respectively, and ideally $T_I = T_{IV} = 1.0$. Sun et al. calculated the tolerance factors using ionic radii, where the ranges of T_I and T_{IV} are 1.0 to 1.22 and 0.84 to 1.04. Therefore, in this work, we initially set tolerance factors ranging from 0.84 to 1.22. After calculating the 675 compounds, the ranges of T_I and T_{IV} are 0.94 to 1.48 and 0.84 to 1.11, respectively. Thus, in the further UL, we set T_I ranges from 0.94 to 1.22 (the red dashed line in Fig. 2a), and T_{IV} ranges from 0.84 to 1.11 (the blue dashed line in Fig. 2a).

Descriptor of SDEPs

According to our previous work and relevant literature reports[21,41–43], we built descriptors from properties of isolated atoms at the I-, II-, IV-, and X-sites. The nine properties of atomic number (Z), group number (g), covalent radius (R_{cov}), Van der Waals radius (R_{vdw}), valence-electron number (N_v), electron affinity (E_{ea}), dipole polarizability (D_p), first ionization energy (E_{ie}), and Pauling electronegativity (X) are considered in this work. For each elemental property (φ), we calculated the minima and maxima of the absolute values of the sums and differences of elemental properties (SDEPs). Succinctly, we introduce the following notations for an elemental

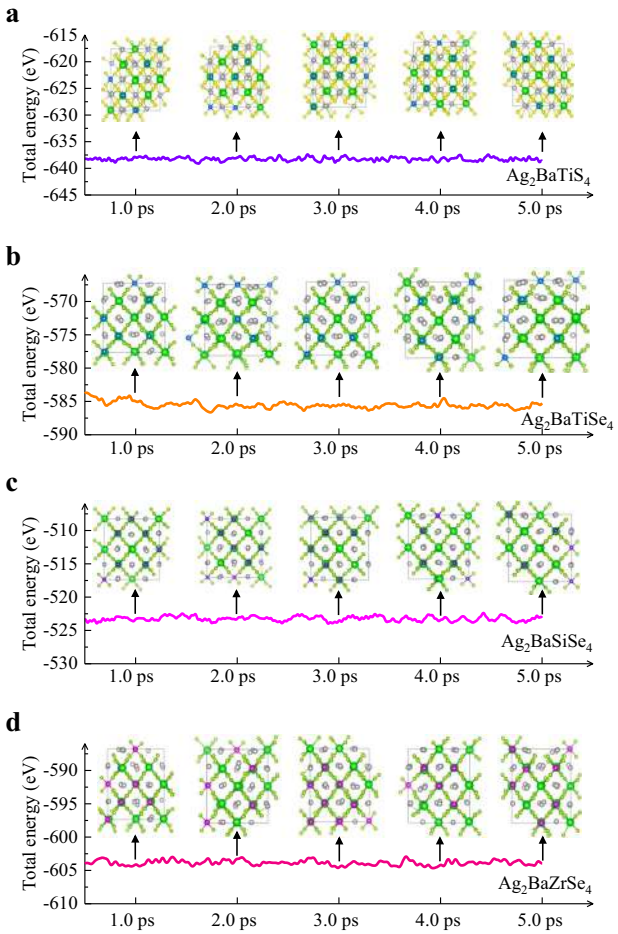

Fig. 7 Evaluations of thermodynamic stabilities of four I_2-II-IV-X_4 chalcogenides by AIMD. a Ag_2BaTiS_4, **b** $Ag_2BaTiSe_4$, **c** $Ag_2BaSiSe_4$, **d** $Ag_2BaZrSe_4$. The thermodynamic stabilities of other four I_2-II-IV-X_4 chalcogenides are provided in Supplementary Fig. 26. The total energy of each step and the crystal structures of each ps are also presented.

property φ:

$$\Delta_{\varphi}^{min\pm} = \min(|\varphi_I \pm \varphi_{II}|, |\varphi_I \pm \varphi_{IV}|, |\varphi_{II} \pm \varphi_{IV}|) \quad (3)$$

$$\Delta_{\varphi}^{max\pm} = \max(|\varphi_I \pm \varphi_{II}|, |\varphi_I \pm \varphi_{IV}|, |\varphi_{II} \pm \varphi_{IV}|) \quad (4)$$

$$\nabla_{\varphi}^{min\pm} = \min(|\varphi_I \pm \varphi_X|, |\varphi_{II} \pm \varphi_X|, |\varphi_{IV} \pm \varphi_X|) \quad (5)$$

$$\nabla_{\varphi}^{max\pm} = \max(|\varphi_I \pm \varphi_X|, |\varphi_{II} \pm \varphi_X|, |\varphi_{IV} \pm \varphi_X|) \quad (6)$$

For each of these equations, we can calculate an 18-dimensional vector based on nine properties. Therefore, a 72-dimensional feature vector can be obtained for each I_2-II-IV-X_4 compound (as shown in Fig. 2b, Supplementary Fig. 4, and Supplementary Table 3). Compared with the 36-dimensional feature vectors obtained from the previous nine independent elemental properties (four site elements), the 72-dimensional feature vectors contain more abundant information, especially the information of differences between elemental properties.

Unsupervised algorithm

UL is accomplished by performing AHC, and the dendrogram function used in AHC is from the *SciPy* package[35]. In AHC, the similarity between samples is calculated by a similarity measure, and each sample is reconnected step by step in order to form nodes. The nodes are organized into a bottom-up tree diagram hierarchy (dendrogram), where the leaf nodes of the tree represent a single sample, and non-leaf nodes are generally obtained by merging similar or close sample sets. The Euclidean distance (L2) between two I_2-II-IV-X_4 compounds was used as the similarity metric, and Ward linkage was used to measure group dissimilarity.

The advantage of AHC is that the partition can be stopped at any time, which means that the number of groups (K) can be adjusted dynamically and directly. In this study, $K = 10$ performs well. More details regarding K can be found in Supplementary Fig. 5 and Supplementary Note 2.

First-principles calculation

All first-principles calculations were conducted using the Vienna Ab initio Simulation Package (VASP)[53]. The Perdew–Burke–Ernzarhof (PBE) generalized gradient approximation (GGA) functionals[54] and project-augmented wave (PAW) atom potentials are employed to perform geometric structure optimizations[55,56]. In this work, for the structures with a space group of *I222*, the experimental Ag_2BaSnS_4 (materials id: mp-555166) served as the starting structure for the cell optimizations; for the structures with a space group of $P3_1$, the experimental Cu_2BaSnS_4 (materials id: mp-17954) served as the starting structure, both of which were obtained from the Materials Project[38]. The cutoff energy for the plane-wave basis was set as 500 eV. The structure optimization process was ended when an energy convergence lower than 10^{-5} eV and atomic force less than 0.05 eV/Å. Further, the HSE06 functional was performed for electronic structure calculations and optical properties[36,37], and the high symmetry points of electrons were obtained from the online tool of the seek-path. More details on the calculations are provided in Supplementary Note 4. Further, to see the effect of lattice constant on the band gap[57], we also optimized eight promising I_2-II-IV-X_4 materials by using the high-level meta-GGA functional (strongly constrained and appropriately normed semilocal density functional, SCAN)[58,59], then calculated the band gaps with HSE06 functional, and found that the differences in structures and band gaps are small (see Supplementary Table 7). This indicates that the convergence criterions (energy and atomic forces) we calculated are reasonable.

The AIMD process was used to evaluate the thermal stability, for structure with a space group of *I222*, a 128-atom $2 \times 2 \times 2$ supercells was built, and for space group of $P3_1$, a 196-atom $2 \times 2 \times 2$ supercell was built. There are enough atoms for phase transition simulations[43,60,61]. With a time step of 1.0 fs, a total of 5 ps of kinetic processes were performed for the structures. It is noted that a longer time scale or larger system size would help to build confidence in stability conclusion but that for now, these cell dimensions will work. During this process, the temperature was controlled at 300 K using the Nosé–Hoover thermostat[62,63].

DATA AVAILABILITY

The starting structures for DFT optimization are available from the Materials Project: https://www.materialsproject.org. The 72-dimensional feature vectors of the 380 I_2-II-IV-X_4 compounds are available from Supplementary Materials.

CODE AVAILABILITY

The codes of AHC model from *SciPy* package used in this work are available at: https://github.com//scipy//scipy.

Received: 9 March 2021; Accepted: 20 July 2021;
Published online: 12 August 2021

REFERENCES

1. Kim, J. Y., Lee, J.-W., Jung, H. S., Shin, H. & Park, N.-G. High-efficiency perovskite solar cells. *Chem. Rev.* **120**, 7867–7918 (2020).
2. Li, H. & Zhang, W. Perovskite tandem solar cells: from fundamentals to commercial deployment. *Chem. Rev.* **120**, 9835–9950 (2020).
3. Xia, X. et al. Photochemical conversion and storage of solar energy. *ACS Energy Lett.* **4**, 405–410 (2019).
4. Yue, Q., Liu, W. & Zhu, X. n-Type molecular photovoltaic materials: design strategies and device applications. *J. Am. Chem. Soc.* **142**, 11613–11628 (2020).
5. Yin, J., Molini, A. & Porporato, A. Impacts of solar intermittency on future photovoltaic reliability. *Nat. Commun.* **11**, 4781 (2020).
6. Kim, B. et al. Cu(In,Ga)$(S,Se)_2$ photocathodes with a grown-In CuxS catalyst for solar water splitting. *ACS Energy Lett.* **4**, 2937–2944 (2019).
7. Chen, C. & Tang, J. Open-circuit voltage loss of antimony chalcogenide solar cells: status, origin, and possible solutions. *ACS Energy Lett.* **5**, 2294–2304 (2020).

8. Wang, W. et al. Device characteristics of CZTSSe thin-film solar cells with 12.6% efficiency. *Adv. Energy Mater.* **4**, 1301465 (2014).
9. Shin, D., Ngaboyamahina, E., Zhou, Y., Glass, J. T. & Mitzi, D. B. Synthesis and characterization of an earth-abundant $Cu_2BaSn(S,Se)_4$ chalcogenide for photoelectrochemical cell application. *J. Phys. Chem. Lett.* **7**, 4554–4561 (2016).
10. Shin, D. et al. Earth-abundant chalcogenide photovoltaic devices with over 5% efficiency based on a $Cu_2BaSn(S,Se)_4$ absorber. *Adv. Mater.* **29**, 1606945 (2017).
11. Zhou, Y. et al. Efficient and stable $Pt/TiO_2/CdS/Cu_2BaSn(S,Se)_4$ photocathode for water electrolysis applications. *ACS Energy Lett.* **3**, 177–183 (2018).
12. Teymur, B., Zhou, Y., Ngaboyamahina, E., Glass, J. T. & Mitzi, D. B. Solution-processed earth-abundant $Cu_2BaSn(S,Se)_4$ solar absorber using a low-toxicity solvent. *Chem. Mater.* **30**, 6116–6123 (2018).
13. Kuo, J. J. et al. Origins of ultralow thermal conductivity in 1-2-1-4 quaternary selenides. *J. Mater. Chem. A* **7**, 2589–2596 (2019).
14. Li, Y. et al. Ultralow thermal conductivity of $BaAg_2SnSe_4$ and the effect of doping by Ga and In. *Mater. Today Phys.* **9**, 100098 (2019).
15. Zhu, T. et al. I_2–II–IV–VI_4 (I = Cu, Ag; II = Sr, Ba; IV = Ge, Sn; VI = S, Se): chalcogenides for thin-film photovoltaics. *Chem. Mater.* **29**, 7868–7879 (2017).
16. Sun, J.-P. et al. Structural tolerance factor approach to defect-resistant I_2-II-IV-X_4 semiconductor design. *Chem. Mater.* **32**, 1636–1649 (2020).
17. Woods-Robinson, R. et al. Wide band gap chalcogenide semiconductors. *Chem. Rev.* **120**, 4007–4055 (2020).
18. Ju, M.-G., Dai, J., Ma, L. & Zeng, X. C. Perovskite chalcogenides with optimal bandgap and desired optical absorption for photovoltaic devices. *Adv. Energy Mater.* **7**, 1700216 (2017).
19. Pang, C. et al. Magnetic properties of semiconducting spinel $CdCr_2S_4$ nanostructured films grown by low-pressure metal–organic chemical vapor deposition. *ACS Appl. Electron. Mater.* **1**, 1424–1432 (2019).
20. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
21. Lu, S. et al. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat. Commun.* **9**, 3405 (2018).
22. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 83 (2019).
23. Chen, C. et al. A critical review of machine learning of energy materials. *Adv. Energy Mater.* **10**, 1903242 (2020).
24. van de Ven, G. M., Siegelmann, H. T. & Tolias, A. S. Brain-inspired replay for continual learning with artificial neural networks. *Nat. Commun.* **11**, 4069 (2020).
25. Chen, C., Zuo, Y., Ye, W., Li, X. & Ong, S. P. Learning properties of ordered and disordered materials from multi-fidelity data. *Nat. Comput. Sci.* **1**, 46–53 (2021).
26. Wu, Y., Duan, H. & Xi, H. Machine learning-driven insights into defects of zirconium metal–organic frameworks for enhanced ethane–ethylene separation. *Chem. Mater.* **32**, 2986–2997 (2020).
27. 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).
28. Torrisi, S. B. et al. Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships. *npj Comput. Mater.* **6**, 109 (2020).
29. Wu, Y., Guo, J., Sun, R. & Min, J. Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells. *npj Comput. Mater.* **6**, 120 (2020).
30. Ding, R. et al. Designing AI-aided analysis and prediction models for nonprecious metal electrocatalyst-based proton-exchange membrane fuel cells. *Angew. Chem. Int. Ed.* **59**, 19175–19183 (2020).
31. Ali, A. et al. Machine learning accelerated recovery of the cubic structure in mixed-cation perovskite thin films. *Chem. Mater.* **32**, 2998–3006 (2020).
32. Park, H. et al. Exploring new approaches towards the formability of mixed-ion perovskites by DFT and machine learning. *Phys. Chem. Chem. Phys.* **21**, 1078–1088 (2019).
33. Ouyang, R. Exploiting ionic radii for rational design of halide perovskites. *Chem. Mater.* **32**, 595–604 (2020).
34. Talapatra, A., Uberuaga, B. P., Stanek, C. R. & Pilania, G. A machine learning approach for the prediction of formability and thermodynamic stability of single and double perovskite oxides. *Chem. Mater.* **33**, 845–858 (2021).
35. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
36. Heyd, J. & Scuseria, G. E. Efficient hybrid density functional calculations in solids: assessment of the Heyd–Scuseria–Ernzerhof screened Coulomb hybrid functional. *J. Chem. Phys.* **121**, 1187–1192 (2004).
37. Garza, A. J. & Scuseria, G. E. Predicting band gaps with hybrid density functionals. *J. Phys. Chem. Lett.* **7**, 4165–4170 (2016).
38. Jain, A. et al. The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
39. Jordan, M. I. & Mitchell, T. M. Machine learning: trends, perspectives, and prospects. *Science* **349**, 255 (2015).
40. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
41. Zhuo, Y., Mansouri Tehrani, A. & Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.* **9**, 1668–1673 (2018).
42. Gladkikh, V. et al. Machine learning for predicting the band gaps of ABX_3 perovskites from elemental properties. *J. Phys. Chem. C.* **124**, 8905–8918 (2020).
43. Wang, Z., Zhang, H. & Li, J. Accelerated discovery of stable spinels in energy systems via machine learning. *Nano Energy* **81**, 105665 (2021).
44. Zhang, Y. et al. Unsupervised discovery of solid-state lithium ion conductors. *Nat. Commun.* **10**, 5260 (2019).
45. Hinuma, Y. et al. Discovery of earth-abundant nitride semiconductors by computational screening and high-pressure synthesis. *Nat. Commun.* **7**, 11962 (2016).
46. Sluydts, M., Pieters, M., Vanhellemont, J., Van Speybroeck, V. & Cottenier, S. High-throughput screening of extrinsic point defect properties in Si and Ge: database and applications. *Chem. Mater.* **29**, 975–984 (2017).
47. Borlido, P. et al. Exchange-correlation functionals for band gaps of solids: benchmark, reparametrization and machine learning. *npj Comput. Mater.* **6**, 96 (2020).
48. Wang, Z. et al. Deep learning for ultra-fast and high precision screening of energy materials. *Energy Storage Mater.* **39**, 45–53 (2021).
49. Zheng, H. et al. Monolayer II-VI semiconductors: a first-principles prediction. *Phys. Rev. B* **92**, 115307 (2015).
50. Torrisi, S. B., Singh, A. K., Montoya, J. H., Biswas, T. & Persson, K. A. Two-dimensional forms of robust CO_2 reduction photocatalysts. *npj 2D Mater. Appl.* **4**, 24 (2020).
51. Park, H. et al. Importance of structural deformation features in the prediction of hybrid perovskite bandgaps. *Comput. Mater. Sci.* **184**, 109858 (2020).
52. Park, H. et al. Data-driven enhancement of cubic phase stability in mixed-cation perovskites. *Mach. Learn. Sci. Technol.* **2**, 025030 (2021).
53. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
54. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
55. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
56. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
57. Singh, A. K., Zhuang, H. L. & Hennig, R. G. Ab initio synthesis of single-layer III-V materials. *Phys. Rev. B* **89**, 245431 (2014).
58. Sun, J., Ruzsinszky, A. & Perdew, J. P. Strongly constrained and appropriately normed semilocal density functional. *Phys. Rev. Lett.* **115**, 036402 (2015).
59. Sun, J. et al. Accurate first-principles structures and energies of diversely bonded systems from an efficient density functional. *Nat. Chem.* **8**, 831–836 (2016).
60. van Gog, H. et al. Thermal stability and electronic and magnetic properties of atomically thin 2D transition metal oxides. *npj 2D Mater. Appl.* **3**, 18 (2019).
61. Lanigan-Atkins, T. et al. Two-dimensional overdamped fluctuations of the soft perovskite lattice in $CsPbBr_3$. *Nat. Mater.* **20**, 977–983 (2021).
62. Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **81**, 511–519 (1984).
63. Hoover, W. G. Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A* **31**, 1695–1697 (1985).

ACKNOWLEDGEMENTS

The work is supported by the National Key Laboratory of Science and Technology on Micro/Nano Fabrication, China, and the National Natural Science Foundation of China (No. 21901157).

AUTHOR CONTRIBUTIONS

Z.W.: Conceptualization, methodology, visualization, data curation, and writing—original draft. J.C.: Methodology, validation, and writing. Q.W. and S.W.: Methodology and validation. J.L.: Conceptualization, methodology, supervision, resources, and writing— review and editing.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-021-00596-4.

Correspondence and requests for materials should be addressed to J.L.