Unsupervised discovery of thin-film photovoltaic materials from unlabeled data

Zhilong Wang1,2, Junfei Cai1,2, Qingxun Wang1,2, SiCheng Wu1,2 and Jinjin Li3✉

Quaternary chalcogenide semiconductors (I2-II-IV-X4) are key materials for thin-film photovoltaics (PVs) to alleviate the energy crisis. Scaling up of PVs requires the discovery of I2-II-IV-X4 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 I2-II-IV-X4 that alleviates the challenge of data scarcity. We screen all the I2-II-IV-X4 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 I2-II-IV-X4 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 energy1,2. 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 sources3-5. Practical thin-film PV cells are based on quaternary chalcogenides (I2-II-IV-X4) of sphalerite crystals such as CdTe and Cu(In, Ga)(S, Se)2 (CIGS), 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 present6,7. However, these materials require expensive or rare elements (In, Te), or even toxic (Cd), severely limiting their large-scale development. Kesterite Cu2ZnSn(S, Se)4 (CZTS) is a potential thin-film PV material, in which the In and Ga in CIGS are replaced with Zn and Sn, and its record PCE of 12.6% is significantly lower than that of CdTe/CIGS8. One possible reason for this is the antisite disorder in the kesterite structure, which significantly affects the open-circuit voltage and device performance. CZTS, where the smaller Zn of is replaced by Ba with a larger ionic radius (Cu2BaSnS4, Se4 (CBTSSe)) to ease the antisite disorder, has demonstrated better performance compared to CZTS in terms of CZTS8-12. Continuing this process, replacing Cu with Ag in CBTSSe to form Ag2BaSnSe4 (ABTSSe), and replacing Ba with Sr to form Cu2SrSnSe4 (CSTSSe), yields materials that have recently shown great promise with respect to thin-film PV applications13,14.

These cases enlighten us that it is worth exploring the wider space of I2-II-IV-X4 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 I2-II-IV-X4 materials is to synthesize or theoretically calculate the properties of massive sets of potential structures in terms of element substitution (e.g., I = Li+, CuII+ or Ag+, II = Ba2+ or Sr2+, IV = Sn4+ or Ge4+; X = O2−, S2−, or Se2−), and then screen for materials with good electro-optical properties. For example, semiconductors with band gaps (Eg) (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 PVs18,19. 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 I2-II-IV-X4 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 systems20-23. 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 104 catalytic samples to design non-noble metal electrocatalytic proton exchange membrane fuel cells30. 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 database31. Moreover, there are some remarkable reports on the application of ML methods in PV materials32-34. 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 I2-II-IV-X4 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.
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 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 of the selected compounds, and successfully discovered eight I$_2$-II-IV-X$_4$ chalcogenides with good electro-optical properties (Ag$_2$BaSnSe$_4$, Ag$_2$BaSnSe$_4$, Ag$_2$BaCrS$_4$, Ag$_2$BaSiSe$_4$, Ag$_2$BaZrS$_4$, Ag$_2$BaZrSe$_4$, Ag$_2$BaHfSe$_4$, and Cu$_2$BaMnSe$_4$). 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), element selection from the periodic table, 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, I$_2$22, P3$_1$, Ama2, and I42m, we selected one structure from each space group as initial structures, as shown in Fig. 1a (see elemental sites and structures in Supplementary Fig. 1). Then, the proper site elements were selected according to the oxidation state and coordination number from the periodic table, where I = Li$^+$, Na$^+$, K$^+$, Cu$^+$, or Ag$^+$; II = Ca$^{2+}$, Sr$^{2+}$, Ba$^{2+}$, Eu$^{2+}$, or...
Pb$^{2+}$; IV= Ti$^{4+}$, Zr$^{4+}$, Hf$^{4+}$, Cr$^{4+}$, Mn$^{4+}$, Co$^{4+}$, S$^{4+}$, Ge$^{4+}$, Sn$^{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$_4$ chalcopyrites 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$_4$ chalcopyrites 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$_4$ chalcopyrites

The process used here for UL is shown in Fig. 1f, consisting of three parts: data set, feature engineering, and algorithm$^{19,40}$. 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, $T_f$ ($T_{1 I}$ and $T_{1 V}$) 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 $0.94 < T_{1 I} < 1.22$ and $0.84 < T_{1 V} < 1.11$ (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$_4$ chalcopyrites 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 $E_g$ 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 1. To construct the feature vectors, we proposed a...
The different colors in Fig. 2c correspond to different groups. More (dendrogram) generated by the AHC is presented in Fig. 2c, compounds are classi where an appropriate partition line is selected and the 380 Cu2BaSnSe4, Cu2BaSnS4, and Ag2BaSnSe4) are shown in Fig. 2b. The feature plots of the six dimensional SDEP in this way for each I2-II-IV-X4 compound (see Supplementary Table 3. Therefore, a visible clustering of I2-II-IV-X4 differentiated, and the SDEPs share similar characteristics within the same groups (Supplementary Fig. 4). Details of the 72-

As the input of clustering, these feature curves look very similar

cant differences (see dotted boxes), such as

can be found using AHC. From Fig. 2c, d, most of the

features of the selected six structures are very similar, but

There are still significant differences (see dotted boxes), such as

The clustering of I2-II-IV-X4 chalcogenides by SDEPs provides

known I2-II-IV-X4 compounds with Eg < 2.0 eV are clustered into G3 in the dendrogram (Fig. 2d), including ten structures of Ag2BaGeSe4 (Eg = 0.85 eV), ABTSe (Eg = 1.42 eV), Cu2BaGeSe4 (Eg = 1.88 eV), and CBTSSe (Eg = 1.96 eV), etc. In addition, G1, including Li2BaGeSe4 (Eg = 2.4 eV) and Li2BaSnSe4 (Eg = 3.07 eV), are all known compounds with Eg > 2.0 eV. For G5 and G8, each contains only one known compound with Eg > 2.0 eV, Li2EuGeSe4 (Eg = 2.54 eV), and Li2PbGeSe4 (Eg = 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 I2-II-IV-X4 chalcogenides. Besides, we also performed the K-means method to cluster I2-II-IV-X4 compounds, the comparison of AHC and K-means are provided in Supplementary Fig. 6, Supplementary Table 4, and Supplementary Note 3.

**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 Eg of 27 known compounds, ten known compounds in G8, and other 17 known compounds not clustered in G8. 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 Eg = 2.0 eV. c Five known I2-II-IV-X4 semiconductors in G3, where I = Ag. d Five known I2-II-IV-X4 semiconductors in G8, where I = Cu.

**Physical insights from unsupervised learning**

The clustering of I2-II-IV-X4 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). G3 and G10 had the most compounds at 52 each, while G2 and G6 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, G3 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 G3 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 G3 deserve further targeted study.

As shown in Fig. 3b, the violin plot of the ten known compounds in G3 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 G3 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 I2-II-IV-X4 compounds with lower $E_g$ among the 26 unknown compounds in G3 with good electronic-optical properties. Moreover, from Supplementary Table 5 and Supplementary Fig. 7, we found that the 36 compounds in G3 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. These discoveries from UL have important guiding significance for the design of I2-II-IV-X4 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 five known compounds with I = Ag$^+$ and II = Ba$^{2+}$: shown in Fig. 3c. These structures contain two kinds of similar space groups: I222 (40%) and I222 (60%), where the I-X4 tetrahedra are flattened and share edges with the II-X4 dodecahedra, do not bear any resemblance to the square antiprisms observed in the P31 and Amma2 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, Amma2 (20%) and P31 (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 P31 is better). To further study the 26 unknown compounds in G3 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 P31. 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 G3, 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 I2-II-IV-X4 compounds (Supplementary Table 3). In addition to speeding up the screening process, UL can estimate important features for guiding the designing of high-performance I2-II-IV-X4 compounds. Therefore, the supervised learning method may be further developed to accurately analyze the importance of features.

### Electro-optical properties of eight I2-II-IV-X4 chalcogenides

We performed ab initio calculations to predict the electro-optical properties of the 26 I2-II-IV-X4 structures in G3. 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

| Compound | Space group | Lattice constant (Å) | Lattice angle (°) | Volume (Å³) | $E_g$ (eV) |
|----------|-------------|---------------------|------------------|-------------|------------|
| Ag$_2$BaTiS$_4$ | I222 | a = 6.50, b = 7.46, c = 8.28 | $\alpha = \beta = \gamma = 90$ | 395.75 | 2.13 |
| Ag$_2$BaTiSe$_4$ | I222 | a = 6.77, b = 7.74, c = 8.54 | $\alpha = \beta = \gamma = 90$ | 447.37 | 0.70 |
| Ag$_2$BaCrS$_4$ | I222 | a = 6.54, b = 7.18, c = 8.16 | $\alpha = \beta = \gamma = 90$ | 383.74 | 1.60 |
| Ag$_2$BaZrS$_4$ | I222 | a = 6.59, b = 7.49, c = 8.41 | $\alpha = \beta = \gamma = 90$ | 415.37 | 2.13 |
| Ag$_2$BaZrSe$_4$ | I222 | a = 6.83, b = 7.86, c = 8.70 | $\alpha = \beta = \gamma = 90$ | 466.91 | 1.76 |
| Ag$_2$BaHfS$_4$ | I222 | a = 6.58, b = 7.49, c = 8.40 | $\alpha = \beta = \gamma = 90$ | 414.09 | 1.33 |
| Ag$_2$BaHfSe$_4$ | I222 | a = 6.82, b = 7.89, c = 8.67 | $\alpha = \beta = \gamma = 90$ | 466.67 | 2.27 |
| Ag$_2$BaSiS$_4$ | I222 | a = 7.02, b = 7.44, c = 8.36 | $\alpha = \beta = \gamma = 90$ | 436.22 | 2.24 |
| Cu$_2$BaTiS$_4$ | P31 | a = b = 6.29, c = 15.73 | $\alpha = \beta = 90, \gamma = 120$ | 538.00 | 2.20 |
| Cu$_2$BaTiSe$_4$ | P31 | a = b = 6.58, c = 16.48 | $\alpha = \beta = 90, \gamma = 120$ | 619.00 | 2.09 |
| Cu$_2$BaCrS$_4$ | P31 | a = b = 6.21, c = 15.32 | $\alpha = \beta = 90, \gamma = 120$ | 511.18 | 2.30 |
| Cu$_2$BaCrSe$_4$ | P31 | a = b = 6.49, c = 16.17 | $\alpha = \beta = 90, \gamma = 120$ | 590.19 | 2.09 |
| Cu$_2$BaMnS$_4$ | P31 | a = b = 6.18, c = 15.29 | $\alpha = \beta = 90, \gamma = 120$ | 505.29 | 0.38 |
| Cu$_2$BaMnSe$_4$ | P31 | a = b = 6.48, c = 15.99 | $\alpha = \beta = 90, \gamma = 120$ | 580.86 | 2.69 |
| Cu$_2$BaZrS$_4$ | P31 | a = b = 6.39, c = 16.00 | $\alpha = \beta = 90, \gamma = 120$ | 565.61 | 2.40 |
| Cu$_2$BaZrSe$_4$ | P31 | a = b = 6.67, c = 16.83 | $\alpha = \beta = 90, \gamma = 120$ | 648.93 | 2.54 |
| Cu$_2$BaCoS$_4$ | P31 | a = b = 6.21, c = 15.00 | $\alpha = \beta = 90, \gamma = 120$ | 500.60 | 3.01 |
| Cu$_2$BaHfS$_4$ | P31 | a = b = 6.36, c = 16.09 | $\alpha = \beta = 90, \gamma = 120$ | 563.16 | 2.73 |
| Cu$_2$BaHfSe$_4$ | P31 | a = b = 6.67, c = 16.82 | $\alpha = \beta = 90, \gamma = 120$ | 647.77 | 3.19 |

Optimized lattice constants, lattice angles, volumes, and band gaps are presented.
D, I direct band gap, indirect 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 \( I = \text{Cu}^+ \). The other five structures have direct band gaps, all from the structures with \( I = \text{Ag}^+ \) (1.42 eV for \( \text{Ag}_2\text{BaTiS}_4 \), 1.18 eV for \( \text{Ag}_2\text{BaTiSe}_4 \), 2.13 eV for \( \text{Ag}_2\text{BaHfSe}_4 \), 1.76 eV for \( \text{Ag}_2\text{BaHfSe}_4 \), and 1.33 eV for \( \text{Ag}_2\text{BaSiSe}_4 \)).

The edges of the conduction band minimums (CBMs) and valence band maximums (VBM}s) of 20 semiconductors are shown in Fig. 4. We found that when \( I = \text{Ag}^+ \), CBMs and VBM{s} are symmetrical 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 \( I = \text{Cu}^+ \), most CBMs and VBM{s} are asymmetric, and CBMs are generally farther from Fermi level than VBM{s}. 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 VBM{s} 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 VBM{s} are mainly influenced by IV-site atoms. All the CBMs and VBM{s} 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 \( G_x \). The above results indicate the importance of elements in regulating CBMs and VBM{s} 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 worthy noting that there are four structures (1.42 eV for \( \text{Ag}_2\text{BaTiS}_4 \), 1.18 eV for \( \text{Ag}_2\text{BaTiSe}_4 \), 1.33 eV for \( \text{Ag}_2\text{BaSiSe}_4 \), and 1.60 eV for \( \text{Ag}_2\text{BaZrSe}_4 \) as shown in Fig. 4) having \( E_g \) between 0.9 and 1.6 eV, which is the range of optimal optical conversion efficiency\(^{10}\). In particular, \( \text{Ag}_2\text{BaTiS}_4 \), \( \text{Ag}_2\text{BaTiSe}_4 \), and \( \text{Ag}_2\text{BaSiSe}_4 \) 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, \( \text{Ag}_2\text{BaCrS}_4 \), \( \text{Ag}_2\text{BaZrS}_4 \), \( \text{Ag}_2\text{BaHfS}_4 \), and \( \text{Cu}_2\text{BaMnSe}_4 \) have both 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 (\( \alpha \)) of these eight structures based on the HSE06 functional are presented in Fig. 6 and Supplementary Fig. 24. They all show strong optical responses (\( \alpha > 10^3 \text{cm}^{-1} \)) 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\(_4\) chalcogenide 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\(^{54,55}\), thus we also calculated the formation enthalpies for eight I\(_2\)-II-IV-X\(_4\) candidates.
candidates with two host space groups of I$\overline{2}22$ and P$\overline{3}$1. As shown in Supplementary Fig. 27, all structures with a space group of I$\overline{2}22$ show the lower formation enthalpies than the structures with space group of P$\overline{3}$1, especially seven structures with I = Ag$^+$ show the lower formation enthalpies than experimental Ag$_2$BaSnS$_4$, indicating their good stability. It is noting that the selected Cu$_2$BaMnSe$_4$ with a space group of P$\overline{3}$1 has a slightly higher formation enthalpy than Cu$_2$BaMnSe$_4$ with a space group of I$\overline{2}22$ and experimental Cu$_2$BaSnS$_4$, 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$\overline{2}$-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$\overline{2}$-II-IV-X$_4$ chalcogenides have become important materials for thin-film PVs. The discovered I$\overline{2}$-II-IV-X$_4$ chalcogenides meet the criteria for earth-abundance and environmental friendliness, and...
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 12-II-IV-X4 compounds, which can be expanded to other material systems. (2) eight I2-II-IV-X4 compounds (Ag2BaTiS4, Ag2BaTiSe4, Ag2BaCrS4, Ag2BaSiSe4, Ag2BaZrS4, Ag2BaZrSe4, Ag2BaHfSe4, and Cu2BaMnSe4) 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 I2-II-IV-X4 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 I2-II-IV-X4 structures, such as the expanded Shannon radii, cell volume33. 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 I2-II-IV-X4 semiconductors, which can be used for structure prediction in an empirically driven learning model15. For I2-II-IV-X4 materials, two dimensionless tolerance factors (TI and TIV) describing the geometric relations have been derived recently. The formula is as follows,

\[ T_I = \sqrt{\frac{4 + \sqrt{2} \lambda_I + \lambda_II}{3} + \lambda_X} \]  
\[ T_{IV} = \sqrt{\frac{4 + \sqrt{2} \lambda_{IV} + \lambda_I}{3} + \lambda_X} \]  

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 TI = TIV = 1.0. Sun et al. calculated the tolerance factors using ionic radii, where the ranges of TI and TIV 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 TI and TIV are 0.94 to 1.48 and 0.84 to 1.11, respectively. Thus, in the further UL, we set TI ranges from 0.94 to 1.22 (the red dashed line in Fig. 2a), and TIV 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 reports21,41–43, we built descriptors from properties of isolated atoms at the I-, II-, IV-, and X-sites. The nine properties of atomic number (N), group number (g), covalent radius (Rcoval), Van der Waals radius (Rvdw), valence-electron number (Nve), electron affinity (Eea), dipole polarizability (Dp), first ionization energy (Ei), and Pauling electronegativity (X) are considered in this work. For each elemental property (gi), 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.

Unsupervised algorithm

UL is accomplished by performing AHC, and the dendrogram function used in AHC is from the SciPy package\(^\text{10}\). In AHC, the similarity between samples is calculated by a similarity measure, and each sample is connected 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 ($L_2$) 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)\(^\text{36}\). The Perdew–Burke–Ernzerhof (PBE) generalized gradient approximation (GGA) functionals\(^\text{54}\) and project-augmented wave (PAW) atom potentials are employed to perform geometric structure optimizations\(^\text{55,56}\). 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\(^\text{36}\). 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 used for electronic structure calculations and optical properties\(^\text{36,57}\), 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\(^\text{55,56}\), 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)\(^\text{36,58}\), 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 $2 \times 2 \times 2$ supercell 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\(^\text{43,60,61}\). 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\(^\text{43,62}\).

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. CuIn(Ga)I$_3$Se$_2$ photocathodes with a grown-In CuS 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).
Correspondence and requests for materials should be addressed to J.L.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021