EES Solar

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d6el00026f

Data-driven discovery of novel chalcogenide semiconductors for solar absorption

Md Habibur Rahman and Arun Mannodi-Kanakkithodi *

The remarkable tunability of chalcogenide semiconductors offers exciting opportunities for applications in solar cells, while also posing significant challenges in exploring their vast compositional space. Although many ternary and quaternary chalcogenides exhibit excellent optoelectronic properties, their performance is often limited by poor defect tolerance and unfavorable doping behavior. Composition engineering is as a promising strategy to simultaneously optimize bulk stability, electronic structure, and defect physics in these materials. In this work, we applied a data-driven framework integrating high-throughput density functional theory (DFT) computations with descriptor-based and structure-based machine learning models to design novel multinary chalcogenide semiconductor alloys ideal for solar absorption and other optoelectronic applications. Within a pre-defined chemical space of zincblende-derived A_2BCX_4 and ABX_2 compounds, we performed hybrid HSE06 calculations with spin–orbit coupling to generate a large dataset of the optimized lattice parameters, decomposition energy, band gap, theoretical maximum PV efficiency, and point defect formation energies for thousands of compounds. Random forest regression models utilizing composition-weighted elemental features were trained and deployed to predict properties for nearly half a million possible compositions, leading to the identification of ~1200 stable compounds with desired optoelectronic properties. Crystal graph-based machine learning force field (MLFF) models were additionally trained on the DFT dataset to enable rapid energy prediction and geometry optimization of new bulk and defect-containing configurations. This workflow led to the identification of several promising compounds, with a notable example being $Cu_2Ca_{0.5}Cd_{0.5}SnS_4$, which satisfies conditions of thermodynamic stability, photovoltaic-suitable band gap, and intrinsic defect tolerance. The entire computational workflow and dataset have been packaged and released as ChalcoDB, an online tool publicly available *via* the nanoHUB platform, thus facilitating community access and ready simulations and predictions for chalcogenide compounds.

Received 11th February 2026
Accepted 22nd April 2026

DOI: 10.1039/d6el00026f

rsc.li/EESSolar

Broader context

The demand for sustainable photovoltaic materials has intensified research into multinary chalcogenide semiconductors, which offer tunable optoelectronic properties and potential for earth-abundant compositions. However, their vast compositional space and complex defect chemistry present significant exploration challenges. Recent advances integrating high-throughput density functional theory (DFT) with machine learning (ML) enable accelerated screening of candidate materials. ML-accelerated DFT provides a great avenue for engineering composition, structure, ionic ordering, and defect behavior in chalcogenide semiconductor alloys. Here, we present a framework powered by high-throughput simulations, descriptor-based regression models, and ML force fields, for discovering novel chalcogenide compounds that are stable, defect-tolerant, and show promising optoelectronic properties. We discuss essential data-driven insights and the prospect of new chalcogenide alloys for solar absorption.

Introduction

Photovoltaic (PV) technology research has intensified in response to rising global demand for sustainable energy.[1–7] CdTe stands out among thin-film absorbers due to its commercial viability, combining strong efficiency with cost-effective production.[2,4,8–10] Yet, environmental hazards from Cd and supply constraints for Te undermine its future prospects.[11] These drawbacks have spurred interest in $Cu(In,Ga)Se_2$ (CIGS) from the ABX_2 chalcogenide class of materials.[12] By adjusting In/Ga ratios, CIGS achieves bandgap optimization for enhanced solar spectrum utilization and efficiency.[13] While less environmentally problematic than CdTe, CIGS faces commercialization hurdles due to In and Ga resource scarcity and price volatility.[14]

Although Cu_2ZnSnS_4 (CZTS) solar cells have been considered a potential alternative, their efficiency still lags behind that of CdTe and CIGS devices. This performance gap is largely attributed to the presence of deep intrinsic point defects that serve as nonradiative recombination centers, reducing both

School of Materials Engineering, Purdue University, West Lafayette, IN 47907, USA. E-mail: amannodi@purdue.edu

minority carrier lifetime and overall device efficiency, something that has been extensively explored and reported in the literature.[15–27] Studies have identified defects such as S vacancy (V_S) and cation anti-site substitution (*e.g.*, Sn_{Zn}) which show low formation energies and often introduce deep defect states detrimental to charge transport.[28] Anion vacancies in CZTS, while sometimes electrically neutral, can still act as problematic defect centers by enabling electron capture and bipolaron formation, where Sn^{4+} is reduced to Sn^{2+}. This stabilization of neutral V_S contributes to recombination losses, lowering the open-circuit voltage (V_{oc}) of CZTS solar cells.[15,28–30]

A study from Minbashi *et al.*[31] demonstrated that controlling defect types and densities in $Cu_2ZnSn(S,Se)_4$ solar cells is crucial for improving PV performance. "Benign" defects with trap densities below 10^{16} cm^{-3} and "harmful" defects above this threshold have been identified from their work, showing that reducing harmful defects led to a record power conversion efficiency (PCE) of 19.0%. Strategies such as Na nanocrystal synthesis, stoichiometry control (*e.g.*, Cu-poor and Zn-rich conditions), and double cation substitution (*e.g.*, Cu with Ag and Zn with Cd) effectively minimized recombination losses and improved device efficiency. Additionally, the incorporation of elements such as Ag and Ga was shown to suppress stacking faults and grain boundaries, which otherwise act as electron traps and degrade device performance.[31]

By comparison, CIGS solar cells demonstrate lower deep-level defect densities, which contributes to their superior efficiencies.[12] Incorporating Se in place of S within CIGS produces a reduced bandgap that suppresses harmful defect formation.[13] Despite having comparable zincblende-based tetrahedral coordination structures, CdTe, CIGS, and CZTS display substantially different defect chemistry.[11] Intrinsic point defects in CZTS and CIGS are further complicated by the emergence of secondary phases during fabrication, including ZnS or Cu_2SnS_3 in CZTS systems, which alter absorber layer electronic behavior.[29] Therefore, controlling secondary phase development and refining chalcogenide absorber stoichiometry are essential for improved material performance.[1,32,33] High-throughput experimental and computational strategies are required to screen and determine optimal compositions and processing parameters that reduce defect levels and enhance optoelectronic characteristics in ternary and quaternary chalcogenides.[9]

Composition engineering *via* alloying at the cation or anion sites plays a critical role in enhancing the efficiency and stability of CZTS-based solar cells, as demonstrated for example in the study on $Cu_2Zn_{1-x}Ba_xSnS_4$ (CZBTS) quinary alloy thin films.[34] Total substitution of Zn with Ba changes the bandgap from 1.48 eV to 1.92 eV. The incorporation of Ba not only induced a structural transition from the kesterite (tetragonal) phase to a trigonal phase, but also improved the crystallinity and reduced defect densities in the films. Ba substitution mitigated cation disorder by avoiding the formation of detrimental Cu–Zn antisite defects, which are known to limit the performance of CZTS solar cells. Overall, the tailored bandgap and enhanced structural properties render CZBTS a promising absorber layer in tandem solar cells, showcasing the potential of composition engineering to optimize properties of chalcogenides for efficient solar energy harvesting.[34]

In recent work from our group,[35] we employed high-throughput density functional theory (DFT) computations to study $A_2^{1+}B^{2+}C^{4+}X_4^{2-}$ quaternary chalcogenides, considering a set of monovalent cations A, divalent cations B, tetravalent cations C, and chalcogen anions X.[35] Similarly, we also simulated $A^{1+}B^{3+}X_2^{2-}$ ternary compounds by considering monovalent cations A, trivalent cations B, and chalcogen anions X.[35] A total of 540 compounds with two types of cation ordering were simulated using the hybrid HSE06 functional with spin–orbit coupling (SOC) to yield their optimized structures, stability, and optoelectronic properties. We identified 45 stable compounds with theoretical maximum PV efficiency exceeding 30%, indicating strong potential as single-junction solar cell absorbers. Further analysis of point defects in two promising candidates revealed susceptibility to harmful anti-site substitutional defects, a known challenge in multinary chalcogenide compounds. It was concluded that further composition engineering through targeted alloying at cation or anion sites is necessary to enable effective multi-objective optimization across a vast combinatorial compositional space and achieve the desired defect tolerance.

Here, we built upon our recent work and applied a combination of high-throughput DFT and different machine learning (ML) approaches to perform multi-objective composition engineering within the ABX_2 and A_2BCX_4 chemical spaces. *Via* controlled mixing at the A, B, C, and X sites, we leveraged a variety of bandgap and stability "bowing"[36–38] effects to achieve optimal thermodynamic stability and optoelectronic properties. For A_2BCX_4 compounds, we considered monovalent cations A $\subset$ {Na, K, Rb, Cs, Cu, Ag}, divalent cations B $\subset$ {Mg, Ca, Sr, Ba, Zn, Cd}, tetravalent cations C $\subset$ {Sn, Ge, Zr}, and chalcogen anions X $\subset$ {S, Se, Te}, allowing fractional occupancies to enable systematic mixing at all cation and anion sites. Through combinatorial analysis, this led to a vast space of 476 280 possible compounds, including 648 pure compounds and diverse mixed compositions across A, B, C, and X sites. Similarly, for the ABX_2 series, we considered monovalent cations A $\subset$ {Na, K, Rb, Cs, Cu, Ag}, trivalent cations B $\subset$ {Al, Ga, In}, and chalcogen anions X $\subset$ {S, Se, Te}, with fractional occupancies leading to 56 700 potential compositions, including 108 pure compounds. All pure compositions adopted two types of cation ordering, namely Kesterite-type and Stannite-type; upon alloying, there are many more possibilities for ionic order and disorder.

By performing DFT computations for a representative subset of this massive chemical space, predictive ML models[39–41] could be trained for screening across all possible compounds at a fraction of the cost of full DFT.[42–53] Such models utilize unique compositional or structural descriptors that uniquely characterize each material and feed them as the input to algorithms such as neural networks or random forest regression to yield the properties of interest as the output. We thus established a "DFT-ML" framework for predicting the structural, electronic, and optical properties of any given pure or alloyed A_2BCX_4 or ABX_2 compounds. During data generation, a series of bulk and defect

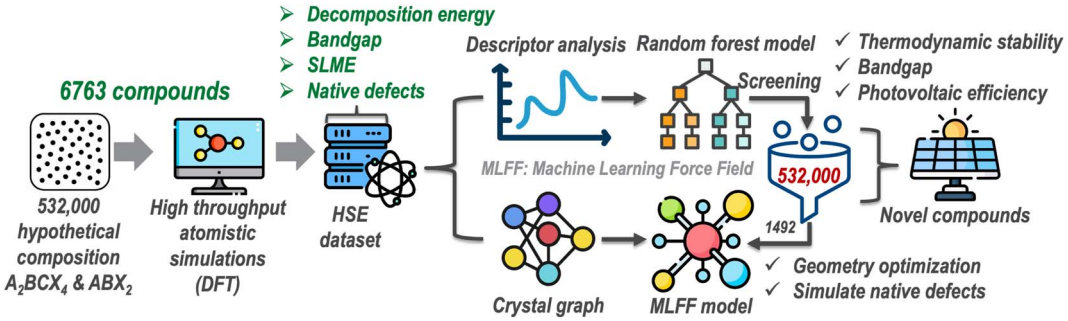

Fig. 1 Computational workflow for discovering novel chalcogenide semiconductors. A large pool of 532 000 hypothetical compositions is down-selected to 6763 candidates for high-throughput DFT computations. The resulting HSE + SOC dataset is used to calculate key properties such as the decomposition energy, bandgap, and photovoltaic efficiency. Descriptor analysis and composition-based random forest models are applied for screening based on stability (decomposition energy < 0 eV), suitable bandgap (1–2 eV), and high photovoltaic efficiency (>30%). Crystal graph neural network-based machine learning force field (MLFF) models are additionally trained on the HSE dataset and employed for geometry optimization and native defect simulations on the most promising compounds identified from ML prediction and screening, leading to the discovery of novel stable and defect-tolerant compounds.

calculations were performed at the HSE06 level of theory, ensuring high training data fidelity and reliable subsequent ML predictions.

Composition-based property predictor models were first trained using random forest regression by employing descriptors derived from the weighted averages of readily available elemental properties of all cation and anion species that made up a given composition. Then, a machine learning force field (MLFF) model based on the Materials 3-body Graph Network (M3GNet)[48] architecture was trained by sampling thousands of structures and the corresponding energies, atomic forces, and stresses from DFT geometry optimization runs, thus achieving the ability to efficiently optimize any new structures of screened compositions. Finally, some of the most promising compounds were studied using DFT accelerated by the MLFF, especially to simulate and understand their defect physics.

All ML training followed the standard best practices in data science, including rigorous hyperparameter optimization, training-test splits, and cross-validation to ensure robust model performance. Once trained, these models enabled rapid screening across thousands of candidates, effectively identifying the compounds with bulk thermodynamic stability, PV-suitable bandgaps, and strong optical absorption characteristics. This integrated DFT-ML framework offers the following key capabilities (Fig. 1):

1. A high-throughput computational workflow to generate an HSE06 + SOC (high-fidelity) dataset of semiconductor properties, along with data visualization tools for extracting qualitative trends.

2. On-demand prediction of the energy of decomposition, electronic bandgap, and theoretical maximum PV efficiency for any A_2BCX_4 or ABX_2 chalcogenide composition (including alloyed variants).

3. An MLFF trained to predict atomic forces and energies at hybrid functional accuracy for both ordered and disordered chalcogenide crystals, including both bulk and defect configurations, enabling efficient geometry optimization *via* gradient-based methods.

4. Accelerated simulations of point defects in the most promising candidates using MLFF-accelerated DFT to evaluate their defect tolerance and dopability.

The overall computational workflow is as follows: 532 000 defined compounds → 6763 representative DFT calculations → thermodynamic, optoelectronic, and defect properties → train RF models and the MLFF → RF screening of 532 000 compounds → select top candidates → MLFF-accelerated DFT defect studies.

Methodology

Chemical space

CZTS and related compounds crystallize in two tetragonal phases, kesterite ($I\bar{4}$) and stannite ($I\bar{4}2m$), which differ in terms of Cu and Zn ionic ordering relative to Sn. The kesterite phase is derived from the chalcopyrite structure of $CuInS_2$ (CIGS), where In is replaced by Zn and Sn, whereas the stannite phase originates from a CuAu-like ordering.[35] In prior work, we investigated the chalcopyrite CIGS structure, referred to as "Ordering 1", and a stannite-type ternary structure denoted as "Ordering 2", where In replaces Zn and Sn in the stannite CZTS. A comprehensive discussion of the effect of these structural orderings on the properties of interest can be found in our previous publication.[35] Because cation alloying blurs the distinction between the prototypical kesterite and stannite frameworks, it is no longer meaningful to strictly label the resulting structures as one or the other. Therefore, for each composition we considered at least two nominal cation orderings, which we denote as Ordering I (kesterite) and Ordering II (stannite) for the A_2BCX_4 family, and similarly as Ordering I and Ordering II for the corresponding ABX_2 variants.

In this work, we considered two large chemical spaces of ternary and quaternary chalcogenide semiconductors. For the quaternary A_2BCX_4 family, the full enumeration of compositions and cation orderings yields 476 280 possible compounds → from this space, we performed DFT calculations on 6146 carefully selected and representative compounds. Likewise, for the ternary ABX_2 family, the total chemical space contains 56

700 compounds → from which 617 representative compounds were computed using DFT. The final DFT dataset consists of 6763 compounds. Importantly, the DFT-calculated subsets are designed to provide a balanced and representative coverage of the full chemical spaces in terms of elemental chemistry, site mixing, and cation ordering. The detailed construction of the chemical spaces and the sampling strategy are provided in the SI.

DFT details

Fig. S1 shows the DFT workflow implemented here, involving geometry optimization, and electronic structure and optical absorption calculations, initially using the semi-local GGA functional followed by the hybrid HSE06 functional. Computations were performed using the Vienna *Ab initio* Simulation Package (VASP) version 6.4.1 with the projector augmented wave (PAW) pseudopotentials.[54,55] The Perdew–Burke–Ernzerhof functional parameterized for solids (PBEsol) within the generalized gradient approximation (GGA) and the hybrid HSE06 functional ($\alpha = 0.25$) were both employed for the exchange–correlation energy.[56–58] A total of 6763 compounds were simulated using 64-atom, $2 \times 2 \times 1$ supercells, across all A_2BCX_4 and ABX_2 pure and alloyed compositions. As an example, to simulate the $Ag_2Ca_{0.5}Sr_{0.5}Sn_{0.5}Ge_{0.5}S_2Te_2$ alloy with Ordering I, we started from the 64-atom kesterite Ag_2ZnSnS_4 structure consisting of 16 Ag, 8 Zn, 8 Sn, and 32 S atoms, and systematically substituted 4 Zn atoms with Ca and another 4 Zn atoms with Sr, 4 Sn atoms with Ge, and half of the S atoms with Te.

Geometry optimization with HSE06 was performed on top of the PBEsol-optimized configurations. Static HSE06 calculations with spin–orbit coupling (henceforth referred to as HSE + SOC) were finally performed on the HSE06-optimized geometries to accurately determine the decomposition energy, electronic bandgap, optical dielectric constant, and optical absorption spectrum of all 6763 compounds. Additionally, we simulated some selected compounds (314 in total) using a larger $3 \times 3 \times 2$ supercell containing 288 atoms using Γ point only. For these calculations, geometry optimization was performed using PBEsol and then using static HSE06, but because of computational expense, no HSE + SOC calculations were performed. These structures were additionally used for performing selected native defect calculations, as will be explained in the following sections. A complete statistical overview of the dataset from different DFT functionals is presented in Table 1. For a comprehensive summary of all computational parameters, the reader is referred to Table S1.

The sequential PBEsol → HSE relaxation protocol ($2 \times 2 \times 2$) is justified by the small geometry differences between the two levels of theory: across all compounds, lattice constants differ by a mean of only 0.27% (std. dev. 0.76%) (see Fig. S2), and internal atomic coordinates show correspondingly small displacements (<0.02 Å), consistent with the known behavior of these functionals for zincblende-derived chalcogenides. As a representative benchmark, detailed comparison of the PBEsol and HSE relaxed structures for $Ag_1Al_{0.5}Ga_{0.5}S_2$ (Table S2) shows that lattice constants agree to within 1% and mean bond lengths differ by 0.03 Å. The slight structural differences also show that ultimately, it is meaningful to perform the higher-fidelity HSE geometry optimization for accurate results, with the initial PBEsol optimization providing notable acceleration. SOC effects on equilibrium geometry are negligible (<0.1% change in lattice parameters) for this class of materials, while the electronic structure corrections are significant. The mean SOC bandgap correction is −0.18 eV, with Te- and Se-containing compounds showing larger corrections (−0.19 eV) than sulfides (−0.15 eV). The static HSE + SOC calculation on the HSE-optimized geometry therefore accurately captures both structural and electronic properties while remaining computationally tractable.

DFT computed properties

The list of properties computed for each compound from HSE + SOC include formation energy (ΔH_{form}), decomposition energy (ΔH_{decomp}), bandgap (E_{gap}), static/optical dielectric constant ($\varepsilon_{optical}$), and the spectroscopic-limited maximum efficiency (SLME) derived from the optical absorption spectrum.[59,60] E_{gap} is obtained from an accurate electronic structure calculation and is one of the most important properties of interest here. Although the $\varepsilon_{optical}$ is not strictly of interest for solar absorption, it is easily extracted from the dielectric function in the optical absorption calculation and thus included here. A large dielectric constant is indeed necessary for wide bandgap semiconductor applications such as power electronics,[61] and the ability to predict it would be important. Furthermore, obtaining the SLME (as a function of semiconductor film thickness) from the optical absorption spectrum is now routine and described in detail in various past publications.[35] We also calculated defect formation energies for selected compounds following the same methodology as described in our earlier work.[35]

ΔH_{form} and ΔH_{decomp} respectively show the energy required for any compound to decompose to elemental species A/B/C/X and to binary A–X, B–X, and C–X phases, and are calculated using the equations below:

$$\Delta H_{form} = \frac{E_{compound} - \sum_i (x_i \times E_i)}{N_{atoms}} \quad (1)$$

$$\Delta H_{decomp} = \frac{E_{compound} - \sum_i (x_i \times E_i) + k_B T \left(\sum_i x_i \ \ln \ x_i \right)}{N_{fu}} \quad (2)$$

In eqn (1), species i refer to the ions at A/B/C/X sites and x_i represents the fractions in which they appear in the compound, whereas in eqn (2), E_i are the energies of all binary phases (A_2X, BX, *etc.*) and x_i account for mixing fractions. The final term in eqn (2) is the mixing entropy in alloyed compositions at a particular temperature T. N_{atoms} is the number of atoms in the compound so that the formation energy is per atom, and N_{fu} is the number of formula units in the compound, where one formula unit is defined as ABX_2 for ternary compounds and $AB_{0.5}C_{0.5}X_2$ for quaternary compounds. As an example, for the

Table 1 Summary of the computational dataset, types of structures (bulk *vs.* defects), and supercell sizes considered in this study. Each snapshot corresponds to a unique atomic configuration obtained from a DFT calculation. The defect dataset includes over 1000 native point defects—vacancies, interstitials, and anti-site substitutions—generated across 314 compounds to ensure diverse chemical and structural representation

Functional	Compounds	Snapshots	System	Supercell	Properties
HSE	6763	∼90 000	Bulk	2 × 2 × 1	ΔH_{form}, ΔH_{decomp}
HSE	314	∼314	Bulk	3 × 3 × 2	ΔH_{form}, ΔH_{decomp}, E_{gap}
HSE + SOC	6763	N/A	Bulk	2 × 2 × 1	ΔH_{form}, ΔH_{decomp}, E_{gap}, $\varepsilon_{optical}$, SLME
HSE	314	961 (q = +2) 970 (q = +1) 981 (q = 0) 971 (q = −1) 965 (q = −2)	Native defects	3 × 3 × 2	$E_D^f(q)$, CTLs (q = +2 to −2)

mixed composition $Ag_2Ca_{0.5}Sr_{0.5}Sn_{0.5}Ge_{0.5}S_2Te_2$, the ΔH_{form} and ΔH_{decomp} will be given by the following (E(X) is the DFT energy per formula unit or per atom of any system X):

$$\Delta H_{form} = (E(Ag_2Ca_{0.5}Sr_{0.5}Sn_{0.5}Ge_{0.5}S_2Te_2) - 2E(Ag) - 0.5E(Ca) - 0.5E(Sr) - 0.5E(Sn) - 0.5E(Ge) - 2E(S) - 2E(Te))/8$$

$$E_{decomp} = E(Ag_2Ca_{0.5}Sr_{0.5}Sn_{0.5}Ge_{0.5}S_2Te_2) - [0.5\,(E(Ag_2S) + E(Ag_2Te)) + 0.25\,(E(CaS) + E(CaTe) + E(SrS) + E(SrTe) + E(SnS_2) + E(SnTe_2) + E(GeS_2) + E(GeTe_2))] + k_BT\,[2 \times (0.5 \ln 0.5) + 8 \times (0.25 \ln 0.25)].$$

Additional details are provided in the SI.

To extend our analysis from stability and optoelectronic properties to defect physics, we selected a total of 314 compounds spanning the A_2BCX_4 and ABX_2 chemistries and generated thousands of possible native point defects in them (namely, vacancies, self-interstitials, and anti-site substitutions), resulting in over 30 000 unique defect configurations. Out of these, approximately 1000 well representative defect structures—comprising 194 vacancies, 152 self-interstitials, and 654 anti-site substitutions—were selected for geometry optimization using PBEsol followed by a static HSE06 calculation. Table S3 lists all the defects that were simulated in different compounds. Defect simulations used a 288-atom 3 × 3 × 2 supercell for any given pure or alloyed compound. Geometry optimization was performed using Gamma-point only, and symmetry-breaking atomic displacements were systematically introduced to avoid energy constraints during relaxation.[62–64] Each point defect was simulated in five distinct charge states (q = +2, +1, 0, −1, −2) to ultimately compute charge-dependent defect formation energy (E^f) and charge transition levels (CTLs). The E^f and CTLs were calculated using the following equations:

$$E^f(D^q,E_F) = E(D^q, A_2BCX_4/ABX_2) - E(A_2BCX_4/ABX_2) + \mu + q(E_F + E_{VBM}) + E_{corr} \quad (3)$$

$$\varepsilon(q_1/q_2) = \frac{E^f(q_1,\ E_F = 0) - E^f(q_2,\ E_F = 0)}{q_2 - q_1} \quad (4)$$

Here, $E(A_2BCX_4/ABX_2)$ is the total energy of the pristine A_2BCX_4 or ABX_2 supercell, and $E(D^q, A_2BCX_4/ABX_2)$ is the energy of the supercell containing a defect D in a charge state q. E_{VBM} is the valence band maximum computed for the bulk compound, E_F is the Fermi level which ranges from the valence to the conduction band edge, μ is the chemical potential of the atom(s) removed or added to create the defect, and E_{corr} is the charge correction energy computed using the Freysoldt method.[65] The slope of the E^f *versus* E_F plot reveals the stable charge state q for any defect, and the transition level $\varepsilon(q_1/q_2)$ denotes the Fermi level where the defect transitions from one stable charge state to another.

To determine the accessible chemical potential ranges for each element during defect calculations, we used the Doped package to construct chemical potential phase diagrams.[62] This method rigorously considers the thermodynamic stability of the target compound against all known competing phases in the corresponding chemical space. For instance, in the case of Cu_2ZnSnS_4, we included Cu_2S, ZnS, SnS_2, Cu, Zn, Sn, and S as competing phases to enforce realistic stability conditions. The Doped package solves the resulting set of linear inequalities to ensure that the chosen chemical potentials maintain phase stability while satisfying the necessary stoichiometry constraints.

Computational dataset

Effect of supercell size

For simulating crystalline semiconductor alloys, it is essential to utilize sufficiently large supercells to perform ionic mixing, such that the effect of order and disorder on the stability and properties could be adequately accounted for. Prior to compiling the entire HSE + SOC dataset, we performed some tests to determine how the structure and properties may change from the 2 × 2 × 1 supercell to the 3 × 3 × 2 supercell for any given composition. To keep the computational expense manageable, this initial analysis used only the GGA-PBEsol functional. As can be seen from Fig. S3, there is a very good match between optimized lattice constants (for a formula unit of the compound), the ΔH_{decomp}, and the E_{gap} from either choice of supercell size. The minor shifts in the values are a result of the additional degrees of freedom in the larger supercell that causes more energy-lowering distortions or types

of ionic ordering. Thus, we proceed with the assumption that the 2 × 2 × 1 supercell is sufficient for predicting properties for a first-level screening, but acknowledge that larger supercells are necessary for taking into account the effects of cation or anion ordering; this will be discussed further when presenting structure-based MLFF models.

A multi-fidelity pre-screening strategy could further reduce the overall computational cost. Analysis of a subset (∼1600 compounds) where both PBEsol and HSE + SOC $\Delta H_{\mathrm{decomp}}$ are available shows that a PBEsol-based filter ($\Delta H_{\mathrm{decomp}}^{\mathrm{PBEsol}} > 0.20$ eV per f.u.) eliminates ∼470 compounds—94% confirmed unstable at HSE + SOC—with a false-negative rate of only 6.2%. A tiered PBEsol → HSE → HSE + SOC approach could reduce the HSE + SOC calculation burden by ∼25–30% in future high-throughput campaigns.

Visualizing the computational dataset

Fig. 2 presents a visualization of the entire HSE + SOC dataset of 6763 compounds in terms of (a) a plot between $\Delta H_{\mathrm{decomp}}$ and E_{gap}, (b) a plot between E_{gap} and $\varepsilon_{\mathrm{optical}}$, and (c) a plot between E_{gap} and SLME. The shape of the E_{gap} *vs.* SLME plot is characteristic of how the PV efficiency of single-junction absorbers varies with the bandgap while constrained by the Shockley–Queisser limit.[59] Also observed is a characteristic inverse relationship between the $\varepsilon_{\mathrm{optical}}$ and E_{gap}, frequently reported for crystalline materials, polymers, and other material classes.[66] The E_{gap}–$\varepsilon_{\mathrm{optical}}$ relationship is due to the fundamental nature of electronic polarization in materials. In general, in compounds with lower E_{gap}, electrons are more easily excited from the valence band to the conduction band under an external electric field, leading to a larger dielectric response. Conversely, in wide E_{gap} materials, the energy required to promote electrons is much higher, reducing their polarizability and thus lowering the $\varepsilon_{\mathrm{optical}}$.

From Fig. 2(a), it is seen that only a minority of all compounds occur in the region of bulk stability ($\Delta H_{\mathrm{decomp}} <$ 0 eV) and PV-suitable bandgaps. We find that out of the 6763 total compounds, 4464 exhibit $\Delta H_{\mathrm{decomp}} > 0$ eV, indicating that they are not stable against decomposition to competing phases. 974 compounds in total (occupying the shaded region in Fig. 2(a)) show both $\Delta H_{\mathrm{decomp}} < 0$ eV and $1.0 < E_{\mathrm{gap}} < 2.0$ eV, which accounts for 15% of the entire DFT dataset. In addition to meeting the E_{gap} and stability conditions, we find that a total of 588 compounds exhibit SLME > 30%. Furthermore, it was ensured during all geometry optimization runs that the crystal structure remains orthogonal ($\alpha \approx \beta \approx \gamma \approx 90°$), especially during alloying where strain effects and symmetry breaking could lead to significant structural distortions even when $\Delta H_{\mathrm{decomp}} < 0$ eV. To address this, we carefully analyzed the lattice parameters and filtered out distorted structures that deviated significantly from orthogonality. After this refinement, we identified 541 stable compounds with the desired bandgaps and SLME > 30%, some of which are listed in Table 2.

Defect properties

In addition to bulk stability and optoelectronic properties, it is vital to understand a material's defect physics to ascertain its suitability for PV and related applications. Fig. 3 shows defect formation energy diagrams constructed for two example compounds using the HSE06 functional: $AgAl_{0.5}Ga_{0.5}Se_2$ in Ordering I and $Ag_2Ca_{0.5}Cd_{0.5}ZrSe_4$ in Ordering II. In total, around 1000 defect configurations were simulated across 314 compounds. In $AgAl_{0.5}Ga_{0.5}Se_2$ (Fig. 3(a)), the Al vacancy (V_{Al}) and Ag at Se anti-site substitution (Ag_{Se}) defects are lowest in energy and respectively exhibit acceptor-type and donor-type (transitioning to neutral around the middle of the bandgap) behavior, with both defects also showing deep transition levels. In contrast, the defect landscape in $Ag_2Ca_{0.5}Cd_{0.5}ZrSe_4$ (Fig. 3(b)) is dominated by donor-type defects such as the Zr interstitial (Zr_i) and Se vacancy (V_{Se}), which show high formation energies across the band gap. Based on the native defect energetics, these compounds respectively

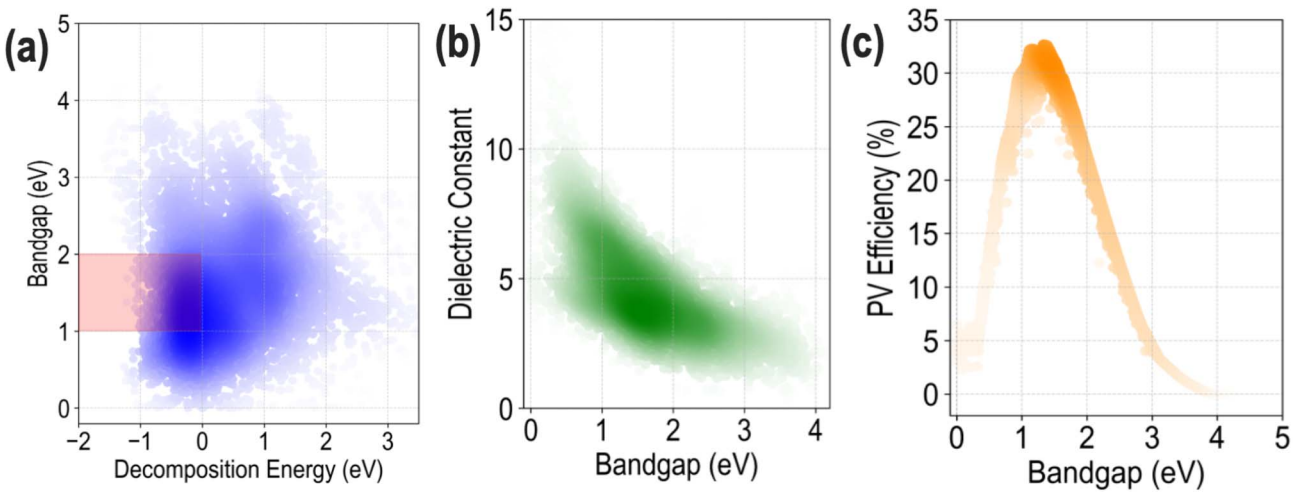

Fig. 2 Visualization of the HSE + SOC dataset: (a) E_{gap} *vs.* $\Delta H_{\mathrm{decomp}}$, (b) $\varepsilon_{\mathrm{optical}}$ *vs.* E_{gap}, and (c) photovoltaic (PV) efficiency *vs.* E_{gap}. Each point is colored by the normalized density estimate, indicating the local data density in the parity plane. Darker regions correspond to areas where many samples cluster, while lighter regions indicate sparsely populated or outlier regions.

Table 2 Selected compounds with ΔH < 0 eV and SLME > 30%

Semiconductor (ordering)	SLME (%)
$Na_{0.5}Cu_{1.5}CdGeSe_4$ (Ordering I)	32.68
$Cu_2ZnGeSSe_3$ (Ordering I)	32.66
$Cu_2MgSnS_2Se_2$ (Ordering I)	32.65
$Ag_2SrSnSe_2Te_2$ (Ordering II)	32.65
$AgAl_{0.5}Ga_{0.5}Te_2$ (Ordering I)	32.65
$Na_{0.5}Ag_{0.5}InTe_2$ (Ordering I)	32.65
$Cu_2CaSnSSe_2Te$ (Ordering II)	32.65
CuAlSeTe (Ordering I)	32.65
$K_{0.5}Cs_{0.5}Cu_{0.5}Ag_{0.5}BaSnTe_4$ (Ordering II)	32.65
$K_{0.5}CuAg_{0.5}ZnSnS_4$ (Ordering II)	32.65
$CuGaTe_2$ (Ordering I)	32.64
$Ag_2CaSnTe_4$ (Ordering II)	32.64
$KCuMgSnSe_4$ (Ordering II)	32.64
$Ag_2MgSnSSe_3$ (Ordering I)	32.64
$Ag_2CaSnS_2Te_2$ (Ordering II)	32.64
$Cu_2CdGeS_2Se_2$ (Ordering I)	32.64
$Cu_{0.5}Ag_{1.5}MgGeSe_4$ (Ordering I)	32.63
$Cu_2Ba_{0.5}Cd_{0.5}SnS_4$ (Ordering I)	32.63
$K_{0.5}Cu_{1.5}ZnSnS_4$ (Ordering II)	32.63
$KCu_{0.5}Ag_{0.5}BaSnTe_4$ (Ordering II)	32.62

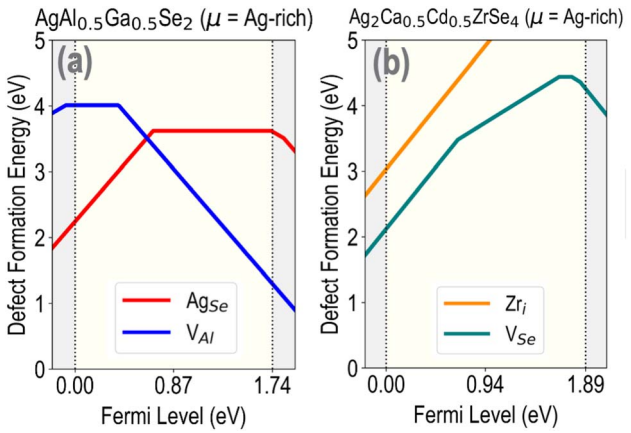

Fig. 3 Defect formation energy (E^f) *vs.* Fermi level (E_F) under Ag-rich conditions for (a) $AgAl_{0.5}Ga_{0.5}Se_2$ in Ordering I, and (b) $Ag_2Ca_{0.5}Cd_{0.5}ZrSe_4$ in Ordering II computed using the HSE06 functional.

show roughly intrinsic and n-type conductivity under Ag-rich conditions, and such information can be similarly gleaned for other materials. While only a couple of compounds are shown here as examples of defect calculations, one of the main purposes of generating the overall defect dataset is to include defect configurations along with bulk structures when training MLFF models, as discussed in the subsequent sections.

Machine learning models

Composition-based regression models

Our ML exercise begins with a smaller initial dataset prior to the expanded set presented in Fig. 2 discussed in the previous section. Using an HSE + SOC dataset of 1650 chalcogenide compounds, we trained three independent random forest regression models to respectively predict ΔH_{decomp}, E_{gap}, and SLME. Each compound is represented by a 50-dimension vector that includes 48 site-weighted averages of elemental properties (ionic radius, electronegativity, oxidation state, and so on) and two dimensions that classify the cation ordering as Ordering I or Ordering II. For ABX_2 compounds, the B site is considered the same as the C site so as to keep the descriptor set definition consistent with A_2BCX_4 compounds. The complete individual set of descriptors for each site is summarized in Table S3 and also discussed in the Pearson correlation subsection in the SI.

The initial random forest training set of 1650 compounds was selected *via* stratified sampling across all cation and anion site occupancies, ensuring balanced coverage of the chemical space.[35] This set included all 540 pure (non-alloyed) compositions plus ~1050 systematically chosen alloys spanning single- and multi-site mixing. Active learning[67] was subsequently used to expand the DFT dataset to 6763 compounds. The MLFF (M3GNet[48]) was trained on snapshots from the HSE geometry optimization trajectories of all 6763 bulk compounds plus defect configurations across 314 compounds, including intermediate relaxation steps to capture the full potential energy surface.

Random forest models[68,69] utilized an 80 : 20 train–test split and 5-fold cross-validation. Grid-based search was used to tune the hyperparameters such as the number of trees, maximum depth, features per split, and the minimum sample counts required for node splitting and leaf formation. This procedure balanced generalization and interpretability while allowing us to extract feature importance rankings from the fitted trees. Parity plots in Fig. S5 compare random forest predictions with the HSE + SOC ground truth for the three properties of interest, also showing uncertainties in prediction based on standard deviations across different trees. Test set root-mean-squared errors (RMSEs) are 0.31 eV for ΔH_{decomp}, 0.25 eV for E_{gap}, and 4.27% for SLME, corresponding to roughly 90–95% accuracy across the observed ranges of the dataset. The larger SLME error is consistent with earlier work[53] that relied solely on composition-based descriptors.

To further broaden the training data in regions where the model exhibited large errors and high prediction uncertainty, we implemented an active learning (AL) loop as illustrated in Fig. S6(a). After each iteration, the random forest ensemble was retrained on the expanded dataset and used to rank unexplored compositions based on an upper-confidence-bound acquisition function, $UCB(x) = \mu(x) + \kappa\sigma(x)$, where $\mu(x)$ is the predicted SLME and $\sigma(x)$ the ensemble variance. The top-ranked candidates were subjected to additional HSE06 + SOC calculations, the new results were incorporated into the dataset for model retraining, and the cycle was repeated. In total, approximately 20 AL rounds contributed an additional 5113 targeted HSE06 + SOC calculations, as pictured in Fig. S6(b), finally resulting in the grand dataset of 6763 HSE + SOC data points presented and discussed earlier. The AL optimization was performed using the SLME value as the objective because of two reasons: (a) among all the models, the SLME predictor initially showed the weakest performance compared to those trained on ΔH_{decomp} and E_{gap},

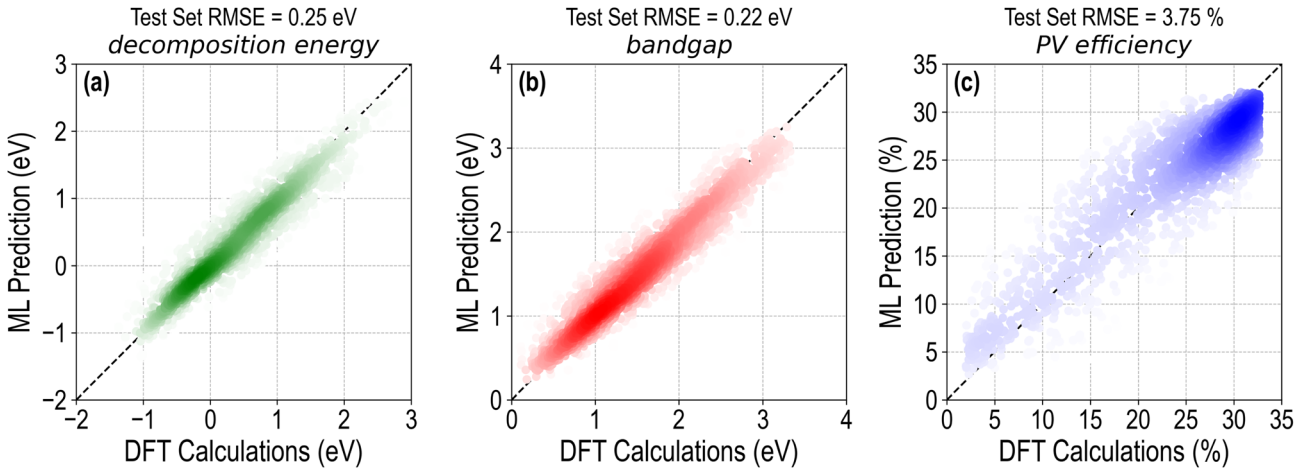

Fig. 4 Test set predictions from random forest regression models trained for different properties: (a) decomposition energy, (b) bandgap, and (c) photovoltaic (PV) efficiency. The test set RMSE values are shown with each parity plot. Darker regions correspond to areas where many test samples cluster, while lighter regions indicate sparsely populated or outlier regions.

and (b) the SLME is ultimately the most important property for designing novel solar absorber materials. Multi-site mixing introduced structural distortions in some compositions, producing outliers that proved problematic for the ML models. To ensure robust training, we removed structures that were heavily distorted after DFT optimization, along with data points exhibiting very high $\Delta H_{\text{decomp}} > 3$ eV per p.f.u., large $E_{\text{gap}} > 4$ eV, and very low SLME (<1%). Incorporating all the new data points reduced the SLME test error to 3.75%, while the test errors for ΔH_{decomp} and E_{gap} fell to 0.25 eV and 0.22 eV, respectively. The final random forest models trained on the expanded dataset are presented in Fig. 4(a–c). The complete curated dataset is provided in the SI.

Structure-based machine learning force field (MLFF) model

We trained a crystal graph neural network (GNN)-based MLFF model, using the M3GNet[48] architecture, on a massive dataset of crystal structures, energies, atomic forces, and stresses sampled from HSE06 geometry optimization calculations of bulk and defect structures across the A_2BCX_4 and ABX_2 chemical space. The optimized MLFF model is being released as part of our online tool, ChalcoDB, for easy use by the community, and is primarily intended to enable rapid geometry optimization of any new bulk or defect configurations. In the M3GNet model, radial and three-body interaction cut-offs were both set to 6 Å.[48,70] The loss function was a weighted RMSE with weights of 1, 1, and 0.01 for energy, forces, and stresses, respectively. Training was carried out from scratch with a batch size of 84 and an initial learning rate of 5×10^{-4}, using a 60 : 20 : 20 train–validation–test split. All training protocols were performed on a single NVIDIA A100 GPU (80 GB Memory).

Fig. 5(a) and (b) present violin plots of the formation-energy distribution, ΔH_{form}, respectively for the entire datasets of bulk and defect configurations used for training the model. Across the total set of 22 000 bulk structures, ΔH_{form} ranges from +100 to −1800 meV per atom, whereas for the set of 1000 defect structures, it spans from −700 to −1500 meV per atom. A few extreme high-energy outliers were removed, but moderately unstable configurations were retained to improve the robustness of the MLFF model. The neutral charge-state ($q = 0$) model was trained on (1) snapshots from the $2 \times 2 \times 1$ supercell bulk dataset; (2) snapshots from the $3 \times 3 \times 2$ supercell bulk dataset;

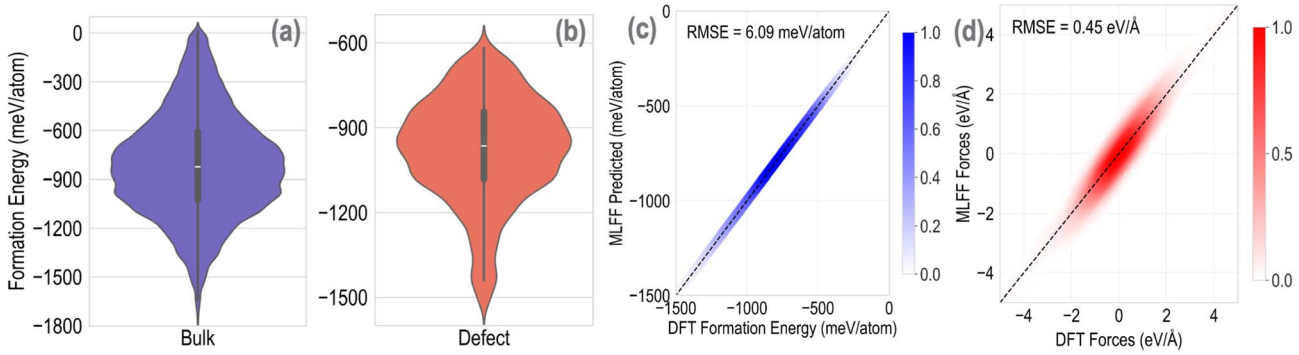

Fig. 5 (a) Violin plot showing DFT-computed formation energies for bulk configurations. (b) Violin plot showing formation energies of defect configurations. (c) MLFF-predicted crystal formation energy compared against DFT-computed values (in meV per atom) across the combined bulk and defect dataset. (d) MLFF-predicted atomic forces compared with DFT values (in eV $Å^{-1}$) across the entire dataset.

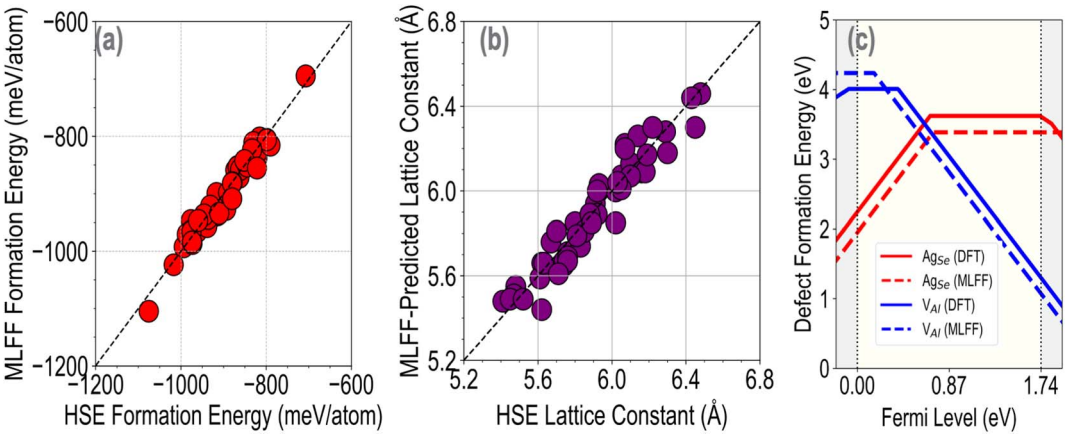

Fig. 6 (a) Parity plot between MLFF-predicted and HSE-computed formation energies of 50 selected test-set compounds. (b) Parity plot between MLFF-predicted and HSE-calculated lattice constants for the same set of 50 compounds, confirming the MLFF accuracy in reproducing ground state structures. (c) Defect formation energy plot comparing MLFF and DFT predictions for two native defects in $AgAl_{0.5}Ga_{0.5}Se_2$ (Ordering I) under Ag-rich conditions.

and (3) snapshots from the 3 × 3 × 2 supercell defect dataset. Fig. 5(c) and (d) show parity plots comparing DFT and MLFF predictions of ΔH_{form} and atomic forces respectively (parity plots for the remaining charge states are provided in the SI). The RMSE corresponding to ΔH_{form} is below 7 meV per atom, while the force RMSE is below 0.45 eV $Å^{-1}$. Using this model, thousands of unexplored chalcogenide compounds can now be optimized with near-HSE accuracy in minutes. Geometry optimizations with the trained MLFF employed the FIRE algorithm in ASE.[71] Following the MLFF-accelerated defect workflow of Mosquera-Lois *et al.*,[70] MLFF geometry optimizations were run for a maximum of 100 ionic steps with a target mean force tolerance of 10^{-5} eV $Å^{-1}$. In practice, all relaxations reached the 100-step limit without satisfying this strict force criterion, with the maximum residual forces converging to approximately 0.01 eV $Å^{-1}$. However, as shown in Fig. S11, the total energy converges within the first 20–30 steps and remains effectively unchanged thereafter (<0.1 meV per atom variation), indicating that the structures are energetically well-converged despite not meeting the formal force threshold.

To handle charged defects, separate M3GNet MLFF models were trained for each charge state (q = +2 to −2), with each model learning the energy-force-stress mapping from DFT data computed at that specific charge state. The charge-state dependence is implicitly captured through the training data rather than an explicit charge encoding in the model architecture. As shown in Fig. S7, all charge-state models achieve comparable accuracy (RMSE of 7–9 meV per atom). Benchmark MLFF relaxation trajectories for representative defects in $Cu_2Ca_{0.5}Cd_{0.5}SnS_4$ (Fig. S11) confirm that the MLFF-relaxed energies converge to within ∼0.5–0.7 eV of DFT reference values across all charge states, with no systematic bias in the sign of the deviation, validating the charge-state-resolved training approach.

To evaluate the accuracy and utility of the MLFF model, we performed the following tests:

1. Energy optimization of bulk structures: we applied the MLFF to optimize the geometries of 50 selected compounds from the test set, and compared the predicted formation energies with the DFT optimized reference energies. Fig. 6(a) shows DFT-optimized *vs.* MLFF-optimized ΔH_{form}, yielding an RMSE of ∼6 meV per atom.

2. Lattice parameter prediction: for the same set of 50 compounds, Fig. 6(b) shows a parity plot for the DFT-optimized and MLFF-optimized lattice constant values, demonstrating agreement within ±1.5% (RMSE <0.1 Å).

3. Computing defect formation energies: Fig. 6(c) presents the defect formation energy diagrams for V_{Al} and Ag_{Se} under Ag-rich conditions in $AgAl_{0.5}Ga_{0.5}Se_2$ (Ordering I). The solid lines correspond to full HSE06 calculations while the dashed lines represent optimized energies from MLFF predictions. The MLFF successfully reproduces the overall E_F dependence and relative energetics of different charge states. This consistency indicates that the MLFF captures the essential defect thermodynamics, including the correct donor or acceptor character reflected in the slope of each segment.

High-throughput prediction and screening

At this stage, we utilized the random forest models to predict ΔH_{decomp}, E_{gap}, and SLME across the entire space of 532 000 possible compounds. Fig. 7(a) presents a visualization of the predicted ΔH_{decomp} *vs.* the E_{gap}, while Fig. 7(b) shows the SLME plotted against E_{gap}. 127 816 compounds were predicted to have $\Delta H_{decomp} < 0$, indicating thermodynamic stability. Among these, 1201 compounds simultaneously exhibit $\Delta H_{decomp} < 0$ and SLME > 30%, some of which are presented in Table 4. Fig. 8(a) shows the screening hierarchy used in this work to identify compounds that satisfy multiple property objectives, culminating in selected defect simulations to evaluate defect tolerance and dopability. An examination of the relative occurrences of different chemical species across the 1201 compounds with optimal properties is pictured in Fig. 8(b).

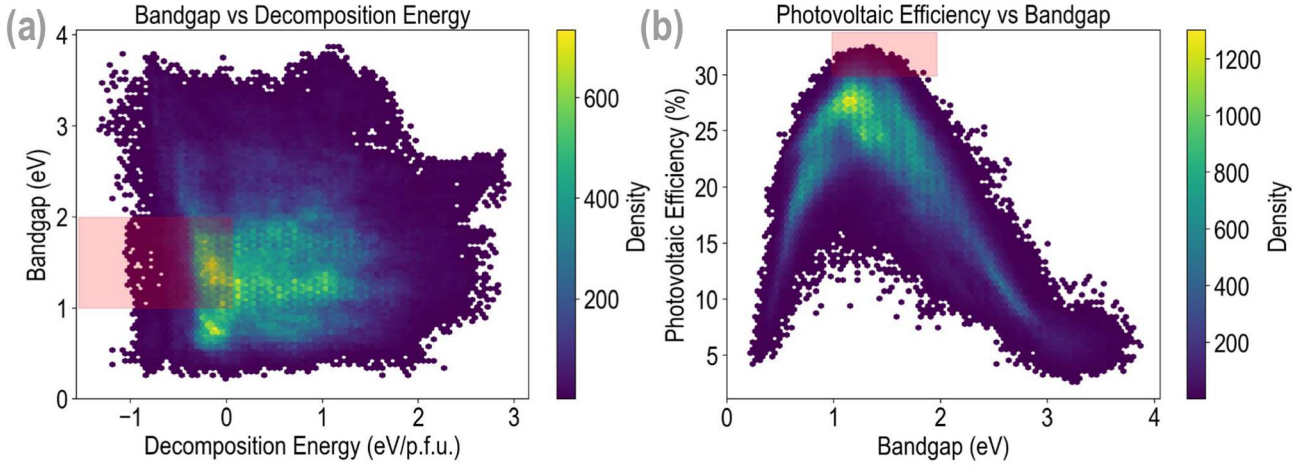

Fig. 7 Visualization of the random forest predictions at the HSE + SOC level across 532 000 compounds: (a) bandgap *vs.* decomposition energy, and (b) photovoltaic (PV) efficiency *vs.* bandgap. Red shading shows the region of interest. Each hexagon represents a bin in two-dimensional property space where the color intensity corresponds to the number (density) of compounds falling within that region. Yellow hexagons indicate higher data density, revealing where candidate materials cluster with similar electronic and thermodynamic properties.

Cu is the dominant A-site cation in both ternary and quaternary systems, while Sn predominates at the C-site in quaternary compounds. Selenides occur more frequently than sulfides or tellurides. The infrequent presence of group I (alkali) and group II (alkaline earth) metals suggests that these elements may be more effective as dopants for property enhancement, particularly for defect tolerance, rather than as primary cation site occupants. Random forest predictions of $\Delta H_{\mathrm{decomp}}$, E_{gap}, and PV efficiency for all 532 000 chalcogenide compounds are provided in the SI as CSV files, along with available DFT (HSE + SOC) data—including lattice parameters, $\Delta H_{\mathrm{decomp}}$, E_{gap}, $\varepsilon_{\mathrm{optical}}$, and optimized structures.

We next used the MLFF models for different charge states to optimize the geometries of native point defects in one of the top candidate compounds, $Cu_2Ca_{0.5}Cd_{0.5}SnS_4$, which shows $E_{\mathrm{gap}}^{\mathrm{DFT}} = 1.32$ eV compared to $E_{\mathrm{gap}}^{\mathrm{ML}} = 1.38$ eV, and $\mathrm{SLME}^{\mathrm{DFT}} = 31.05\%$ compared to $\mathrm{SLME}^{\mathrm{ML}} = 30.92\%$. To generate defect configurations, we used an automated workflow based on the ShakeNBreak protocol[64] to introduce systematic bond distortions. The Doped package was used to automatically determine the chemical potential limits for each defect using the CompetingPhasesAnalyzer module. All native point defects (vacancies, interstitials, and antisites) were structurally optimized across five charge states (−2, −1, 0, +1, and +2) using the MLFF models. Subsequently, single-point HSE06 calculations were performed on the MLFF-optimized structures to obtain accurate final defect formation energies.

Fig. 9(a) shows the optimized crystal structure of $Cu_2Ca_{0.5}Cd_{0.5}SnS_4$ and Fig. 9(b) shows the calculated formation energies for the lowest-energy native defects in this compound under S-rich conditions. V_{Cu} exhibits the lowest formation energy across most of the bandgap region, indicating that a Cu vacancy is the dominant intrinsic acceptor, consistent with the p-type conductivity commonly observed in Cu-based chalcogenides. The V_{Ca} and V_{Cd} defects show higher formation energies closer to the valence band. Among donor-type defects, the Ca_{Cu} cation anti-site substitution defect has the lowest formation energy ($E_f \approx 0.8$ eV at the VBM), and does not show any charge

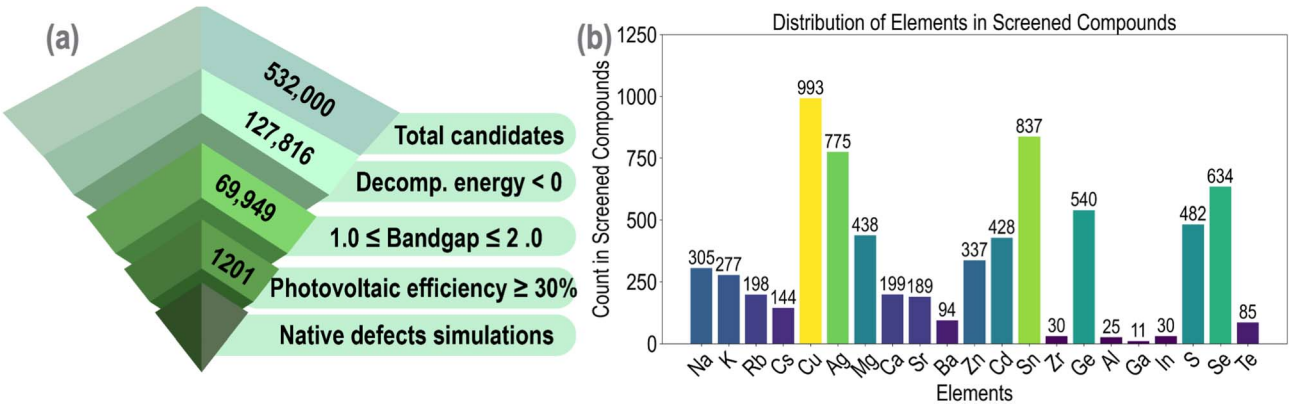

Fig. 8 (a) Screening hierarchy used in this study to discover stable compounds with optimal optoelectronic properties. (b) Distribution of various cations and anions across the set of 1201 promising compounds identified in this work.

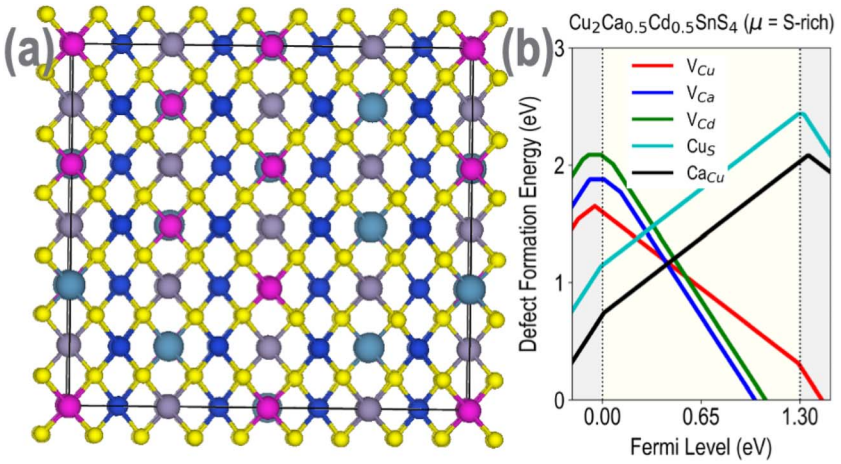

Fig. 9 (a) Optimized crystal structure of $Cu_2Ca_{0.5}Cd_{0.5}SnS_4$. (b) Defect formation energies as a function of Fermi level, computed for $Cu_2Ca_{0.5}Cd_{0.5}SnS_4$ under S-rich chemical potential conditions.

transition level inside the bandgap region. The dominance of low-energy V_{Cu} acceptors and the absence of shallow donor defects suggest that $Cu_2Ca_{0.5}Cd_{0.5}SnS_4$ will exhibit moderately p-type conductivity, with n-type doping likely challenging to achieve.

Since the MLFF-optimized geometries serve as input for final HSE + SOC single-point calculations, consistency between the MLFF potential energy surface and the HSE reference is critical. PBEsol-trained models would yield relaxed geometries that may not correspond to minima on the HSE surface, particularly for defect structures where exchange-driven electron localization effects are significant. To directly illustrate this, Fig. S9 presents parity plots of MLFF-predicted *vs.* DFT-computed atomic forces for both PBEsol- and HSE-trained models on neutral defect configurations. The HSE-trained MLFF achieves a force RMSE of 0.030 eV $Å^{-1}$, lower than that of the PBEsol-trained model (RMSE = 0.051 eV $Å^{-1}$), demonstrating that the HSE potential energy surface is more accurately captured by the MLFF and yields superior force predictions for defect geometry optimization.

While detailed defect formation energy diagrams are presented only for selected case studies for the sake of brevity, the entire defect dataset spans over 1000 native point defect configurations across 314 compounds, each computed using a 3 × 3 × 2 supercell with PBEsol geometry optimization followed by static HSE06 calculations. Representative defect formation energy diagrams for 18 randomly selected compounds under cation-rich chemical potential conditions are provided in Fig. S10. General trends include (i) Cu/Ag vacancies are consistently the lowest-energy acceptors, supporting p-type conductivity; (ii) anion-related defect energetics are sensitive to the anion species, with Te-containing compounds showing higher formation energy defects; (iii) anti-site defects involving B/C-site cations most frequently introduce deep levels in the E_{gap}; and (iv) several compounds exhibit high minimum defect formation energies (>3 eV) across the entire E_{gap}, indicating strong defect tolerance.

To complement the optoelectronic screening, we computed electronic band structures from HSE + SOC for four selected top-ranked candidates: $Cu_2ZnGeSSe_3$, $Cu_2MgSnS_2Se_2$, $AgAl_{0.5}Ga_{0.5}Te_2$, and $Cu_2Ca_{0.5}Cd_{0.5}SnS_4$ (Fig. S8). These compounds span Cu- and Ag-containing systems with diverse anion chemistries and represent candidates with favorable E_{gap} (1.3–1.5 eV), negative ΔH_{decomp}, and high SLME values. The band structures reveal dispersive band edges near the Fermi level, suggesting moderate effective masses conducive to efficient carrier transport. Table 3 presents the computed electron and hole effective masses of these compounds alongside other properties.

Notably, the hole effective masses of all four candidates (0.113–0.162 m_0) are significantly lighter than those of established photovoltaic absorbers such as CdTe($m_h^* \approx 0.69\ m_0$)[72] and Cu_2ZnSnS_4 (CZTS, $m_h^* \approx 0.71\ m_0$).[73] These low hole effective masses indicate highly dispersive valence band edges, which are favorable for efficient hole transport and reduced

Table 3 Key properties of the four selected compounds (all in Ordering II) for which band structures are computed. All properties are computed at the HSE + SOC level of theory. Effective masses are in units of m_0

Compound	E_{gap} (eV)	SLME (%)	ΔH_{decomp} (eV per f.u.)	m_e^* (m_0)	m_h^* (m_0)
$Cu_2ZnGeSSe_3$	1.345	32.66	−0.988	0.917	0.131
$Cu_2MgSnS_2Se_2$	1.332	32.65	−0.765	0.963	0.148
$AgAl_{0.5}Ga_{0.5}Te_2$	1.365	32.65	−0.317	0.480	0.113
$Cu_2Ca_{0.5}Cd_{0.5}SnS_4$	1.522	31.05	−0.362	1.626	0.162

Table 4 Examples of ML-identified compounds with $\Delta H^{ML}_{decomp} < 0$ eV and $SLME^{ML} > 30\%$

Semiconductor (ordering)	E^{ML}_{gap} (eV)	$SLME^{ML}$ (%)
$Cu_2ZnGeSSe_3$ (Ordering II)	1.21	32.32
$KCu_{0.5}Ag_{0.5}MgGeSe_4$ (Ordering I)	1.22	32.26
$Ag_2MgSnSSe_3$ (Ordering II)	1.31	32.26
$Cu_2Ba_{0.5}Cd_{0.5}SnS_4$ (Ordering II)	1.40	32.17
$K_{0.5}Cu_{1.5}ZnSnS_4$ (Ordering I)	1.34	32.15
$Ag_2SrSnSe_2Te_2$ (Ordering I)	1.28	32.11
$K_{0.5}Cu_{0.5}AgZnSnS_4$ (Ordering I)	1.32	32.02
$KCuMgSnSe_4$ (Ordering I)	1.26	32.00
$K_{0.5}Cu_{1.5}Mg_{0.5}Cd_{0.5}SnS_4$ (Ordering I)	1.32	31.96
$K_{0.5}Cu_{0.5}AgZn_{0.5}Cd_{0.5}SnS_4$ (Ordering I)	1.29	31.94
$Cu_2ZnGeS_2Se_2$ (Ordering II)	1.44	31.93
$Cu_2CdGeS_2Se_2$ (Ordering II)	1.37	31.93
$Ag_2Sr_{0.5}Cd_{0.5}SnS_4$ (Ordering I)	1.40	31.92
$K_{0.5}Rb_{0.5}Cu_{0.5}Ag_{0.5}MgGeSe_4$ (Ordering I)	1.17	31.92
$K_{0.5}Cu_{0.5}AgCdSnS_4$ (Ordering I)	1.27	31.91
$Cu_2Zn_{0.5}Cd_{0.5}SnS_4$ (Ordering II)	1.23	31.90
$RbCu_{0.5}Ag_{0.5}MgGeSe_4$ (Ordering I)	1.20	31.90
$KCu_{0.5}Ag_{0.5}MgSnSe_4$ (Ordering I)	1.16	31.89
$K_{0.5}Cu_{0.5}AgMg_{0.5}Cd_{0.5}SnS_4$ (Ordering I)	1.34	31.89
$Cu_2Sr_{0.5}Cd_{0.5}SnS_4$ (Ordering II)	1.40	31.87

recombination losses in photovoltaic devices. Among the four compounds, $AgAl_{0.5}Ga_{0.5}Te_2$ exhibits the lightest hole mass (0.113 m_0), while $Cu_2Ca_{0.5}Cd_{0.5}SnS_4$ has the heaviest (0.162 m_0), yet both remain substantially lighter than CdTe and CZTS.

While the 1201 candidates identified in this work satisfy thermodynamic stability criteria ($\Delta H_{decomp} < 0$ eV) computed against relevant competing binary phases, we acknowledge that experimental synthesis feasibility depends on additional factors not captured by ground-state DFT calculations alone, including kinetic barriers to nucleation and growth, competition with metastable polymorphs, and non-equilibrium processing conditions. It has been realized that configurational mixing entropy (mean contribution of −0.032 eV) stabilizes mixed compositions further, highlighting the thermodynamic favorability of cation mixing in these structures. Many base compounds in our chemical space (*e.g.*, CZTS,[74,75] $CuInS_2$, and $CuGaSe_2$ (ref. 12)) have been experimentally realized, and the alloyed compositions we predict are derived from these *via* systematic cation/anion substitution, which is a well-established experimental strategy.[24,76] In a related recent publication,[77] we developed models to predict the "synthesizability" of perovskite-type compounds by combining experimental and computational data, and found that ΔH_{decomp} has a very clear correlation with the synthesis likelihood, thus reinforcing our belief that a negative ΔH_{decomp} value is a good indicator of thermodynamic feasibility and compound formability. Future work will include developing similar synthesis prediction models for multi-nary chalcogenides as well driving collaborative synthesis and characterization of the most promising candidates to validate our computational predictions. We anticipate that the publicly available ChalcoDB platform and dataset will help guide and prioritize such experimental efforts.

ChalcoDB: an interactive nanoHUB tool

We developed ChalcoDB (https://nanohub.org/tools/chalcodb, Fig. 10) as an open-source interactive web-based platform deployed on nanoHUB for high-throughput computational analysis of A_2BCX_4 (I–II–IV–VI kesterite/stannite) and ABX_2 (I–III–VI chalcopyrite) semiconductors. This platform integrates DFT data with ML models and automated simulation workflows, providing the community with immediate access to geometry optimization, property prediction, defect analysis, and molecular dynamics (MD) simulation capabilities without requiring local computational infrastructure or programming expertise.

DFT database and composition-based ML predictions

The platform provides access to our comprehensive HSE + SOC dataset containing formation energies, ΔH_{decomp}, E_{gap}, $\varepsilon_{optical}$, SLME, and optimized crystal structures for 6763 chalcogenide compounds. Users can explore this database through interactive filtering and visualization tools. For rapid screening of novel compositions, the platform employs random forest regression models trained on this DFT dataset to predict ΔH_{decomp}, E_{gap}, and SLME for user-defined compositions. The composition-based ML models enable immediate property estimates for compounds not yet synthesized or computed, with automatic validation against the DFT database when exact matches exist.

Structure optimization and property prediction

This module enables users to build custom compositions by selecting elements for each crystallographic site with automatic stoichiometry validation. The tool employs the M3GNet MLFF models trained on both the PBEsol and HSE06 datasets to optimize multiple geometries in parallel. Users specify the compound type, starting structure (kesterite, stannite, or chalcopyrite), and supercell dimensions. Each optimized configuration is automatically evaluated in terms of the formation energy, ΔH_{decomp}, E_{gap}, and SLME using M3GNet regression models trained on the ChalcoDB HSE + SOC dataset. The results are presented through interactive tables identifying the lowest-energy configuration, energy convergence plots, and 3D structure visualization with downloadable CIF files.

Defect formation and optimization

Using optimized ground-state structures, this module facilitates the creation and relaxation of point defects including vacancies, substitutional defects, interstitials, and defect complexes. The tool employs MLFF-based geometry optimization to determine equilibrium defect structures across multiple charge states. Users can define defects using intuitive syntax and obtain relaxed structures with energy analysis.

Molecular dynamics simulations

MLFF-based MD[78–83,83–85] capabilities enable users to study finite-temperature behavior, thermal stability, and phase transitions in chalcogenide semiconductors. The MD module

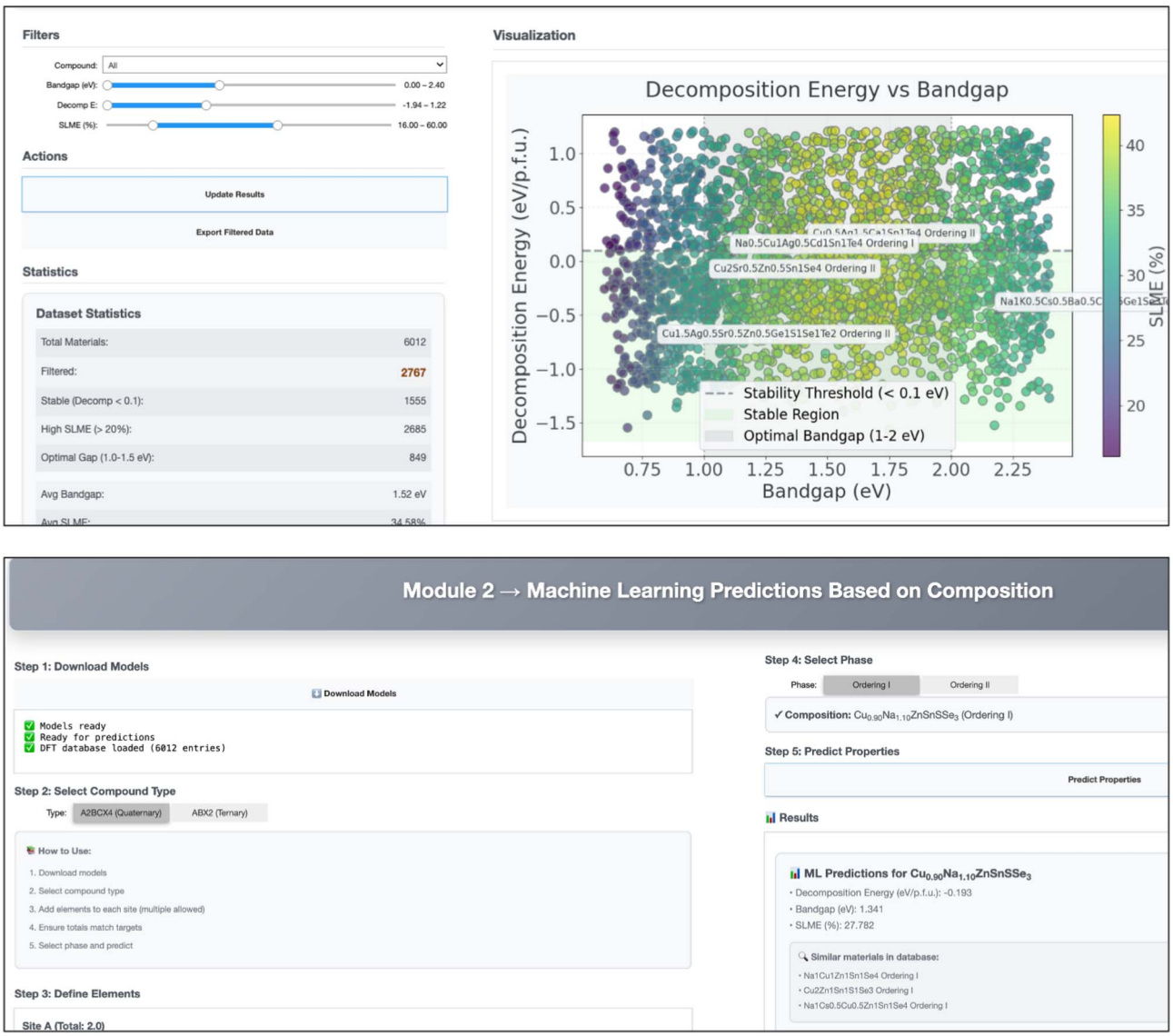

Fig. 10 ChalcoDB – an interactive data and ML platform for computational design of chalcogenide semiconductors. Top: visualization module showing ΔH_{decomp} *versus* E_{gap} for thousands of compounds. Filters enable rapid screening based on ΔH_{decomp}, E_{gap}, and theoretical maximum PV efficiency. Shaded regions indicate ΔH_{decomp} and optimal E_{gap} windows. Bottom: composition-based prediction module where users define A_2BCX_4 or ABX_2 chemistries, select the cation ordering type, and obtain ML predictions of ΔH_{decomp}, E_{gap}, and SLME in real time.

supports NVT, NPT, and NVE[86–91] ensembles with user-defined temperature. Simulations provide trajectories for analysis of structural evolution over time.

Defect migration and diffusion analysis

This module calculates migration barriers for ionic and defect diffusion using the Nudged Elastic Band (NEB) method. The tool automatically identifies potential migration pathways through neighbor analysis (nearest-neighbor, second-nearest-neighbor, or exhaustive site search) and constructs the minimum energy path between the initial and final defect positions. Users specify the defect type, charge state, and migration target, with real-time monitoring of force convergence.

Conclusions

This work establishes a scalable data-driven framework for the discovery and design of defect-tolerant I–II–IV–VI and I–III–VI semiconductors by tightly integrating high-throughput DFT, ML models, and automated screening workflows. By systematically exploring the vast compositional space of zincblende-derived A_2BCX_4 and ABX_2 compounds and associated alloys, we demonstrated that composition engineering combined with predictive modeling can simultaneously optimize thermodynamic stability, optoelectronic properties, and defect tolerance. The identification of over a thousand stable compounds with attractive properties, *e.g.* $Cu_2Ca_{0.5}Cd_{0.5}SnS_4$, highlights the effectiveness of combining descriptor-based models with graph-based machine-learned force fields to accelerate DFT

calculations. The public release of the ChalcoDB platform further enables reproducible and community-driven exploration of chalcogenide semiconductors. Future work will extend this framework to incorporate finite-temperature stability, defect migration and carrier transport, and grain-boundary/interface effects that critically influence device performance. Integrating active learning strategies, uncertainty quantification, and experimental validation will be essential for refining predictions and guiding synthesis. Closed-loop computational–experimental workflows will help realize autonomous discovery and optimization of next-generation chalcogenide semiconductors for high-efficiency PV and optoelectronic technologies.

Conflicts of interest

There are no conflicts to declare.

Data availability

All DFT and ML data are available as part of the supplementary information (SI) and in the nanoHUB tool: https://nanohub.org/tools/chalcodb. All associated code and scripts are available on GitHub: https://github.com/msehabibur/ChalcoDB. Supplementary information: a description of the chemical space, DFT workflow and details, comparisons between properties from different DFT functionals and supercell sizes, regression models with uncertainties and active learning workflow, MLFF models for different charge states, selected electronic band structures, comparison of PBEsol-MLFF and HSE-MLFF models, selected defect formation energy diagrams, selected MLFF geometry optimization plots, description of computed energetics, a list of specific defects that were simulated, and a list of all descriptors used for ML models along with correlation values with different properties. See DOI: https://doi.org/10.1039/d6el00026f.

Acknowledgements

A. M. K. acknowledges support from the Purdue School of Materials Engineering faculty start-up grant and the Defense Advanced Research Projects Agency (DARPA) Young Faculty Award 2025. This material is additionally based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Solar Energy Technology Office (SETO) Award Number DE-0009332. Funding for this work was also provided by the Alliance for Sustainable Energy, LLC, Managing and Operating Contractor for the National Renewable Energy Laboratory for the U.S. DOE, and was supported in part by EERE under SETO Award Number 37989. This work utilized the Anvil cluster at Purdue through allocation MAT230030 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by the U.S. National Science Foundation grant 2138259, 2138286, 2138307, 2137603, and 2138296.

References

1 M. A. Green, E. D. Dunlop, D. H. Levi, J. Hohl-Ebinger, M. Yoshita and A. W. Ho-Baillie, Solar cell efficiency tables (version 54), *Prog. Photovoltaics Res. Appl.*, 2019, **27**, 565–575.
2 S. Rojsatien, A. Mannodi-Kanakkithodi, T. Walker, N. Mohan Kumar, T. Nietzold, E. Colegrove, D. Mao, M. E. Stuckelberger, B. Lai, Z. Cai, M. K. Y. Chan and M. I. Bertoni, Distribution of Copper States, Phases, and Defects across the Depth of a Cu-Doped CdTe Solar Cell, *Chem. Mater.*, 2023, **35**, 9935–9944.
3 C. Li, J. Poplawsky, Y. Yan and S. J. Pennycook, Understanding individual defects in CdTe thin-film solar cells *via* STEM: From atomic structure to electrical activity, *Mater. Sci. Semicond. Process.*, 2017, **65**, 64–76.
4 P. Gorai, D. Krasikov, S. Grover, G. Xiong, W. K. Metzger and V. Stevanović, A search for new back contacts for CdTe solar cells, *Sci. Adv.*, 2023, **9**, eade3761.
5 M. H. Rahman, M. Jubair, M. Z. Rahaman, M. S. Ahasan, K. Ostrikov and M. Roknuzzaman, RbSnX3 (X = Cl, Br, I): promising lead-free metal halide perovskites for photovoltaics and optoelectronics, *RSC Adv.*, 2022, **12**, 7497–7505.
6 M. H. Rahman, M. Z. Rahaman, E. H. Chowdhury, M. Motalab, A. K. M. A. Hossain and M. Roknuzzaman, Understanding the role of rare-earth metal doping on the electronic structure and optical characteristics of ZnO, *Mol. Syst. Des. Eng.*, 2022, **7**, 1516–1528.
7 E. H. Chowdhury, M. H. Rahman, R. Jayan and M. M. Islam, Atomistic investigation on the mechanical properties and failure behavior of zinc-blende cadmium selenide (CdSe) nanowire, *Comput. Mater. Sci.*, 2020, **186**, 110001.
8 C. Ferekides and J. Britt, CdTe solar cells with efficiencies over 15%, *Sol. Energy Mater. Sol. Cells*, 1994, **35**, 255–262.
9 M. H. Rahman, P. Gollapalli, P. Manganaris, S. K. Yadav, G. Pilania, B. DeCost, K. Choudhary and A. Mannodi-Kanakkithodi, Accelerating defect predictions in semiconductors using graph neural networks, *APL Mach. Learn.*, 2024, **2**, 016122.
10 M. Gloeckler, I. Sankin and Z. Zhao, CdTe Solar Cells at the Threshold to 20% Efficiency, *IEEE J. Photovoltaics*, 2013, **3**, 1389–1393.
11 V. Fthenakis, Sustainability of photovoltaics: The case for thin-film solar cells, *Renew. Sustain. Energy Rev.*, 2009, **13**, 2746–2750.
12 P. Jackson, D. Hariskos, E. Lotter, S. Paetel, R. Wuerz, R. Menner, W. Wischmann and M. Powalla, New world record efficiency for Cu(In,Ga)Se2 thin-film solar cells beyond 20%, *Prog. Photovoltaics Res. Appl.*, 2011, **19**, 894–897.
13 M. A. Green, The path to 25% silicon solar cell efficiency: History of silicon cell evolution, *Prog. Photovoltaics Res. Appl.*, 2009, **17**, 183–189.
14 W. Liu, H. Li, B. Qiao, S. Zhao, Z. Xu and D. Song, Highly efficient CIGS solar cells based on a new CIGS bandgap

gradient design characterized by numerical simulation, *Sol. Energy*, 2022, **233**, 337–344.
15 S. Chen, J.-H. Yang, X. G. Gong, A. Walsh and S.-H. Wei, Intrinsic point defects and complexes in the quaternary kesterite semiconductor Cu 2 ZnSnS 4, *Phys. Rev. B*, 2010, **81**, 245204.
16 W. Chen, D. Dahliah, G.-M. Rignanese and G. Hautier, Origin of the low conversion efficiency in Cu2ZnSnS4 kesterite solar cells: the actual role of cation disorder, *Energy Environ. Sci.*, 2021, **14**, 3567–3578.
17 K. Rudisch, A. Davydova, L. Riekehr, J. Adolfsson, L. Q. Casal, C. Platzer-Björkman and J. Scragg, Prospects for defect engineering in Cu2ZnSnS4 solar absorber films, *J. Mater. Chem. A*, 2020, **8**, 15864–15874.
18 R. B. Wexler, G. S. Gautam and E. A. Carter, Optimizing kesterite solar cells from Cu2ZnSnS4 to Cu2CdGe(S,Se)4, *J. Mater. Chem. A*, 2021, **9**, 9882–9897.
19 S. Kim, J.-S. Park, S. N. Hood and A. Walsh, Lone-pair effect on carrier capture in Cu2ZnSnS4solar cells, *J. Mater. Chem. A*, 2019, **7**, 2686–2693.
20 K. Zhao, H. Xiang, R. Zhu, C. Liu and Y. Jia, Passivation principle of deep-level defects: a study of SnZn defects in kesterites for high-efficient solar cells, *J. Mater. Chem. A*, 2022, **10**, 2849–2855.
21 T. Ratz, N. D. Nguyen, G. Brammertz, B. Vermang and J.-Y. Raty, Relevance of Ge incorporation to control the physical behaviour of point defects in kesterite, *J. Mater. Chem. A*, 2022, **10**, 4355–4365.
22 R. Scaffidi, G. Brammertz, Y. Wang, A. U. Zaman, K. Sasikumar, J. deWild, D. Flandre and B. Vermang, A study of bandgap-graded CZTGSe kesterite thin films for solar cell applications, *Energy Adv.*, 2023, **2**, 1626–1633.
23 M. Kauk-Kuusik, K. Timmo, M. Pilvet, K. Muska, M. Danilson, J. Krustok, R. Josepson, V. Mikli and M. Grossberg-Kuusk, Cu2ZnSnS4 monograin layer solar cells for flexible photovoltaic applications, *J. Mater. Chem. A*, 2023, **11**, 23640–23652.
24 J. Guo, J. Ao and Y. Zhang, A critical review on rational composition engineering in kesterite photovoltaic devices: self-regulation and mutual synergy, *J. Mater. Chem. A*, 2023, **11**, 16494–16518.
25 H. Xu, X. Guo, H. Yang, Q. Zhou, S. Liu, H. Gao, C. Gao and W. Yu, Improving the crystallization and properties of CZTSSe film by adding NaTFSI in the precursor solution, *J. Mater. Chem. C*, 2023, **11**, 5498–5504.
26 Y. Zhao, C. Xu, Z. Zhou, Y. Chen, Y. Zhang, L. Wu, X. Su, X. Hu and S. Wang, Enhancing the efficiency of Cu2ZnSn(S,Se)4solar cells by variable-temperature sulfoselenization, *J. Mater. Chem. C*, 2023, **11**, 10660–10672.
27 A. Mannodi-Kanakkithodi, The devil is in the defects, *Nat. Phys.*, 2023, **19**, 1243–1244.
28 S. Kim, J. S. Park and A. Walsh, Identification of Killer Defects in Kesterite Thin-Film Solar Cells, *ACS Energy Lett.*, 2018, **3**, 496–500.
29 J. J. Scragg, J. T. Wätjen, M. Edoff, T. Ericson, T. Kubart and C. Platzer-Björkman, A Detrimental Reaction at the Molybdenum Back Contact in Cu2ZnSn(S,Se)4 Thin-Film Solar Cells, *J. Am. Chem. Soc.*, 2012, **134**, 19330–19333.
30 K. Park, B.-H. Jeong, H. Y. Lim and J.-S. Park, Effect of chemical substitution on polytypes and extended defects in chalcopyrites: A density functional theory study, *J. Appl. Phys.*, 2021, **129**, 025703.
31 M. Minbashi, A. Ghobadi, E. Yazdani, A. A. Kordbacheh and A. Hajjiah, Efficiency enhancement of CZTSSe solar cells *via* screening the absorber layer by examining of different possible defects, *Sci. Rep.*, 2020, **10**, 21813.
32 B. Prakash, A. Meena, Y. K. Saini, S. Mahich, A. Singh, S. Kumari, C. S. P. Tripathi and B. L. Choudhary, Solution-processed CZTS thin films and its simulation study for solar cell applications with ZnTe as the buffer layer, *Environ. Sci. Pollut. Res.*, 2022, **30**, 98671–98681.
33 Y. Li, H. Wei, C. Cui, X. Wang, Z. Shao, S. Pang and G. Cui, CZTSSe solar cells: insights into interface engineering, *J. Mater. Chem. A*, 2023, **11**, 4836–4849.
34 S. K. M, S. P. Madhusudanan, A. R. Rajamani, M. Siaj and S. K. Batabyal, Barium substitution in Kesterite CU2ZNSNS4: CU2ZN1–XBAxSNS4 Quinary Alloy thin films for efficient solar energy harvesting, *Cryst. Growth Des.*, 2020, **20**, 4387–4394.
35 M. H. Rahman and A. Mannodi-Kanakkithodi, High-throughput screening of ternary and quaternary chalcogenide semiconductors for photovoltaics, *Comput. Mater. Sci.*, 2025, **249**, 113654.
36 A. Mannodi-Kanakkithodi, A first principles investigation of ternary and quaternary II–VI zincblende semiconductor alloys, *Model. Simulat. Mater. Sci. Eng.*, 2022, **30**, 044001.
37 A. Mannodi-Kanakkithodi and M. K. Y. Chan, Data-driven design of novel halide perovskite alloys, *Energy Environ. Sci.*, 2022, **15**, 1930–1949.
38 J. Yang, P. Manganaris and A. Mannodi-Kanakkithodi, A high-throughput computational dataset of halide perovskite alloys, *Digital Discovery*, 2023, **2**, 856–870.
39 M. H. Rahman, M. Biswas and A. Mannodi-Kanakkithodi, DeFecT-FF: Accelerated Modeling of Defects in Cd-Zn–Te-Se-S Compounds Combining High-Throughput DFT and Machine Learning Force Fields, *Phys. Chem. Chem. Phys.*, 2026, 00170j.
40 M. H. Rahman, I. Agrawal andA. Mannodi-Kanakkithodi, Using Machine Learning to Explore Defect Configurations in Cd/Zn-Se/Te Compounds, *2025 IEEE 53rd Photovoltaic Specialists Conference (PVSC)*, 2025, pp 0717–0719.
41 M. Tenorio, M. H. Rahman, A. Mannodi-Kanakkithodi and J. Chapman, Out-of-distribution machine learning for materials discovery: Challenges and opportunities, *Chem. Phys. Rev.*, 2026, **7**, 011317.
42 J. Cheng, C. Zhang and L. Dong, A geometric-information-enhanced crystal graph network for predicting properties of materials, *Commun. Mater.*, 2021, **2**, 92.
43 M. Kilgour, J. Rogal and M. Tuckerman, Geometric Deep Learning for Molecular Crystal Structure Prediction, *J. Chem. Theory Comput.*, 2023, **19**, 4743–4756.

44 K. Choudhary and B. DeCost, Author Correction: Atomistic Line Graph Neural Network for improved materials property predictions, *npj Comput. Mater.*, 2022, **8**, 221.
45 T. Xie and J. C. Grossman, Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties, *Phys. Rev. Lett.*, 2018, **120**, 145301.
46 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals, *Chem. Mater.*, 2019, **31**, 3564–3572.
47 A. N. Rubungo, C. Arnold, B. P. Rand and A. B. Dieng, LLM-Prop: predicting the properties of crystalline materials using large language models, *npj Comput. Mater.*, 2025, **11**, 186.
48 C. Chen and S. P. Ong, A universal graph deep learning interatomic potential for the periodic table, *Nat. Comput. Sci.*, 2022, **2**, 718–728.
49 M. H. Rahman, M. Biswas and A. Mannodi-Kanakkithodi, Understanding Defect-Mediated Ion Migration in Semiconductors using Atomistic Simulations and Machine Learning, *ACS Mater. Au*, 2024, **4**, 557–573.
50 A. Mannodi-Kanakkithodi and M. K. Y. Chan, Accelerated screening of functional atomic impurities in halide perovskites using high-throughput computations and machine learning, *J. Mater. Sci.*, 2022, **57**, 10736–10754.
51 A. Mannodi-Kanakkithodi, M. Y. Toriyama, F. G. Sen, M. J. Davis, R. F. Klie and M. K. Y. Chan, Machine-learned impurity level prediction for semiconductors: the example of Cd-based chalcogenides, *npj Comput. Mater.*, 2020, **6**, 39.
52 A. Mannodi-Kanakkithodi, X. Xiang, L. Jacoby, R. Biegaj, S. T. Dunham, D. R. Gamelin and M. K. Y. Chan, Universal machine learning framework for defect predictions in zinc blende semiconductors, *Patterns*, 2022, **3**, 100450.
53 J. Yang, P. Manganaris and A. Mannodi-Kanakkithodi, Discovering novel halide perovskite alloys using multi-fidelity machine learning and genetic algorithm, *J. Chem. Phys.*, 2024, **160**, 064114.
54 P. E. Blöchl, Projector augmented-wave method, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1994, **50**, 17953–17979.
55 G. Kresse and J. Furthmüller, Efficient iterative schemes for ab initiototal-energy calculations using a plane-wave basis set, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 11169–11186.
56 J. P. Perdew, K. Burke and M. Ernzerhof, Generalized Gradient Approximation Made Simple, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.
57 J. Heyd, G. E. Scuseria and M. Ernzerhof, Hybrid functionals based on a screened Coulomb potential, *J. Chem. Phys.*, 2003, **118**, 8207–8215.
58 M. Y. Toriyama, J. Qu, G. J. Snyder and P. Gorai, Defect chemistry and doping of BiCuSeO, *J. Mater. Chem. A*, 2021, **9**, 20685–20694.
59 M. Bercx, N. Sarmadian, R. Saniz, B. Partoens and D. Lamoen, First-principles analysis of the spectroscopic limited maximum efficiency of photovoltaic absorber layers for CuAu-like chalcogenides and silicon, *Phys. Chem. Chem. Phys.*, 2016, **18**, 20542–20549.
60 V. Wang, N. Xu, J.-C. Liu, G. Tang and W.-T. Geng, VASPKIT: A user-friendly interface facilitating high-throughput computing and analysis using VASP code, *Comput. Phys. Commun.*, 2021, **267**, 108033.
61 Y. Zhang, D. Dong, Q. Li, R. Zhang, F. Udrea and H. Wang, Wide-bandgap semiconductors and power electronics as pathways to carbon neutrality, *Nat. Rev. Electr. Eng.*, 2025, **2**, 155–172.
62 S. R. Kavanagh, A. G. Squires, A. Nicolson, I. Mosquera-Lois, A. M. Ganose, B. Zhu, K. Brlec, A. Walsh and D. O. Scanlon, doped: Python toolkit for robust and repeatable charged defect supercell calculations, *J. Open Source Softw.*, 2024, **9**, 6433.
63 I. Mosquera-Lois, S. R. Kavanagh, A. Walsh and D. O. Scanlon, Identifying the ground state structures of point defects in solids, *npj Comput. Mater.*, 2023, **9**, 25.
64 I. Mosquera-Lois, S. R. Kavanagh, A. Walsh and D. O. Scanlon, ShakeNBreak: Navigating the defect configurational landscape, *J. Open Source Softw.*, 2022, **7**, 4817.
65 C. Freysoldt, B. Grabowski, T. Hickel, J. Neugebauer, G. Kresse, A. Janotti and C. G. Van De Walle, First-principles calculations for point defects in solids, *Rev. Mod. Phys.*, 2014, **86**, 253–305.
66 A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman and R. Ramprasad, Machine Learning Strategy for Accelerated Design of Polymer dielectrics, *Sci. Rep.*, 2016, **6**, 20952.
67 D. E. Farache, J. C. Verduzco, Z. D. McClure, S. Desai and A. Strachan, Active learning and molecular dynamics simulations to find high melting temperature alloys, *Comput. Mater. Sci.*, 2022, **209**, 111386.
68 M. H. Rahman and A. Mannodi-Kanakkithodi, Defect modeling in semiconductors: the role of first principles simulations and machine learning, *J. Phys.: Mater.*, 2025, **8**, 022001.
69 F. Pedregosa, *et al.*, Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
70 I. Mosquera-Lois, S. R. Kavanagh, A. M. Ganose and A. Walsh, Machine-learning structural reconstructions for accelerated point defect calculations, *npj Comput. Mater.*, 2024, **10**, 121.
71 A. H. Larsen, *et al.*, The atomic simulation environment—a Python library for working with atoms, *J. Phys. Condens. Matter*, 2017, **29**, 273002.
72 L. S. Dang, G. Neu and R. Romestain, Optical detection of cyclotron resonance of electron and holes in CdTe, *Solid State Commun.*, 1982, **44**, 1187–1190.
73 C. Persson, Electronic and optical properties of Cu_2ZnSnS_4 and $Cu_2ZnSnSe_4$, *J. Appl. Phys.*, 2010, **107**, 053710.
74 W. Wang, M. T. Winkler, O. Gunawan, T. Gokmen, T. K. Todorov, Y. Zhu and D. B. Mitzi, Device Characteristics of CZTSSe Thin-Film Solar Cells with 12.6% Efficiency, *Adv. Energy Mater.*, 2014, **4**, 1301465.
75 C. Yan, *et al.*, Cu2ZnSnS4 solar cells with over 10% power conversion efficiency enabled by heterojunction heat treatment, *Nat. Energy*, 2018, **3**, 764–772.

76 A. S. Nazligul, M. Wang and K. L. Choy, Recent Development in Earth-Abundant Kesterite Materials and Their Applications, *Sustainability*, 2020, **12**, 5138.
77 R. Desai, J. Ahn, A. Strachan and A. Mannodi-Kanakkithodi, Bridging the synthesizability gap in perovskites by combining computations, literature data, and PU learning, *Mach. Learn.: Sci. Technol.*, 2025, **6**, 045061.
78 M. H. Rahman, E. H. Chowdhury and M. M. Islam, Understanding mechanical properties and failure mechanism of germanium-silicon alloy at nanoscale, *J. Nanopart. Res.*, 2020, **22**, 311.
79 M. H. Rahman, E. H. Chowdhury, M. R. B. Shahadat and M. M. Islam, Engineered defects to modulate the phonon thermal conductivity of Silicene: A nonequilibrium molecular dynamics study, *Comput. Mater. Sci.*, 2021, **191**, 110338.
80 E. H. Chowdhury, M. H. Rahman, S. Fatema and M. M. Islam, Investigation of the mechanical properties and fracture mechanisms of graphene/WSe2 vertical heterostructure: A molecular dynamics study, *Comput. Mater. Sci.*, 2021, **188**, 110231.
81 M. H. Rahman, S. Mitra, M. Motalab and P. Bose, Investigation on the mechanical properties and fracture phenomenon of silicon doped graphene by molecular dynamics simulation, *RSC Adv.*, 2020, **10**, 31318–31332.
82 M. H. Rahman, E. H. Chowdhury and S. Hong, High temperature oxidation of monolayer MoS2 and its effect on mechanical properties: A ReaxFF molecular dynamics study, *Surf. Interfaces*, 2021, **26**, 101371.
83 M. H. Rahman, M. S. Islam, M. S. Islam, E. H. Chowdhury, P. Bose, R. Jayan and M. M. Islam, Phonon thermal conductivity of the stanene/hBN van der Waals heterostructure, *Phys. Chem. Chem. Phys.*, 2021, **23**, 11028–11038.
84 M. H. Rahman, E. H. Chowdhury and S. Hong, Atomic-level investigation on the oxidation efficiency and corrosion resistance of lithium enhanced by the addition of two dimensional materials, *RSC Adv.*, 2022, **12**, 5458–5465.
85 M. H. Rahman, S. Mitra, M. Motalab and T. Rakib, Investigation on the temperature and size dependent mechanical properties and failure behavior of zinc blende (ZB) gallium nitride (GaN) semiconducting nanowire, *2020 IEEE Region 10 Symposium (TENSYMP)*. 2020, pp 22–25.
86 M. H. Rahman, E. H. Chowdhury and S. Hong, Nature of creep deformation in nanocrystalline cupronickel alloy: A Molecular Dynamics study, *Results Mater.*, 2021, **10**, 100191.
87 E. H. Chowdhury, M. H. Rahman and S. Hong, Tensile strength and fracture mechanics of two-dimensional nanocrystalline silicon carbide, *Comput. Mater. Sci.*, 2021, **197**, 110580.
88 M. H. Rahman, E. H. Chowdhury, D. A. Redwan, S. Mitra and S. Hong, Characterization of the mechanical properties of van der Waals heterostructures of stanene adsorbed on graphene, hexagonal boron–nitride and silicon carbide, *Phys. Chem. Chem. Phys.*, 2021, **23**, 5244–5253.
89 S. Mitra, M. H. Rahman, M. Motalab, T. Rakib and P. Bose, Tuning the mechanical properties of functionally graded nickel and aluminium alloy at the nanoscale, *RSC Adv.*, 2021, **11**, 30705–30718.
90 E. H. Chowdhury, M. H. Rahman, P. Bose, R. Jayan and M. M. Islam, Atomic-scale analysis of the physical strength and phonon transport mechanisms of monolayer β-bismuthene, *Phys. Chem. Chem. Phys.*, 2020, **22**, 28238–28255.
91 M. H. Rahman, E. H. Chowdhury, D. A. Redwan and S. Hong, Computational characterization of thermal and mechanical properties of single and bilayer germanene nanoribbon, *Comput. Mater. Sci.*, 2021, **190**, 110272.