PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2026, **28**, 10718

DeFecT-FF: a machine learning force field framework for high throughput defect modeling in CdTe-based solar cells

Md Habibur Rahman, Maitreyo Biswas and Arun Mannodi-Kanakkithodi *

We developed a framework for predicting the energies and ground state configurations of native point defects, extrinsic dopants and impurities, and defect complexes across zinc blende-phase Cd/Zn–Te/Se/S compounds, important for CdTe-based solar cells. This framework, named DeFecT-FF, is powered by high-throughput density functional theory (DFT) computations and crystal graph-based machine learning force field (MLFF) models trained on the DFT data. The Cd/Zn–Te/Se/S chemical space is chosen because alloying at Cd or Te sites is a promising avenue to tailor the electronic and defect properties of the CdTe absorber layer to potentially improve solar cell performance. The sheer number of defect configurations achievable when considering all possible singular defects and their combinations, symmetry-breaking operations, and defect charge states, as well as the expense of running large supercell calculations, makes this an ideal problem for developing accurate and widely-applicable force field models. Here, we introduce our dataset of structures and energies from HSE06 geometry optimization, including bulk and alloyed supercells with and without defects. Data were gradually expanded using active learning and accurate MLFF models were trained to predict energies and atomic forces across different charge states. *Via* accelerated prediction and screening, we identified many new low energy defect configurations and obtained high-fidelity defect formation energy diagrams using HSE06 calculations with spin–orbit coupling. The DeFecT-FF framework has been released publicly as an online tool on the nanoHUB platform, allowing users to upload any crystallographic information file, generate defects of interest, and compute defect formation energies as a function of Fermi level and chemical potential conditions, thus bypassing expensive DFT calculations.

Received 17th January 2026,
Accepted 7th April 2026

DOI: 10.1039/d6cp00170j

rsc.li/pccp

1 Introduction

Advancements in solar cell technologies are crucial for meeting global energy demands and supporting the transition to a decarbonized grid.[1–4] Among photovoltaic (PV) technologies, CdTe ranks second after crystalline Si, accounting for about 7% of the global market.[1,2,5] Its commercial success arises from a direct band gap of ~1.5 eV, high absorption coefficient ($>5 \times 10^5$ cm^{-1}),[6–15] low production cost, and good thin-film conductivity.[16] However, its maximum efficiency of 22.3% remains below the ~30% theoretical limit, mainly due to Shockley–Read–Hall (SRH) recombination associated with grain boundaries, point defects, and dislocations.[17] Native defects and impurities create trap states within the band gap that act as nonradiative recombination centers,[18,19] such as Cd vacancies (V_{Cd}), which accelerate carrier recombination and can reduce power conversion efficiency by nearly 5%.[20–24]

CdTe often suffers from low hole density, limiting its PV efficiency.[1,2] Cu is commonly introduced as an acceptor dopant *via* high-temperature $CdCl_2$ annealing, where Cu and Cl diffuse at 10^{17}–10^{19} cm^{-3}, altering electronic properties.[1,2,25] While Cu_i and Cl_{Te} act as shallow donors and Cu_{Cd} as a non-shallow acceptor forming complexes such as (Cu_i + Cu_{Cd}) and (Cl_i + Cu_{Cd})$^{2+}$,[5,26,27] Cu doping typically yields suboptimal hole density (~10^{14} cm^{-3}) compared to the ideal value of ~10^{16} cm^{-3}.[16] In contrast, group V dopants such as As achieve higher hole densities without reducing carrier lifetime.[5,28] Se alloying to create $CdSe_xTe_{1-x}$ further enhances PV efficiency by improving absorption, band alignment, and carrier lifetimes.[28,29] ZnTe, with favorable band alignment, serves as an efficient hole transport layer.[30] Thus, exploring the defect chemistry of Cd/Zn–S/Se/Te alloys is vital to improving CdTe- and CdSeTe-based thin-film solar cells.[31]

In semiconductors, point defects can trap or release electrons, and thus they tend to exist in multiple charge states q depending on the Fermi level position (E_F) within the band gap. For each defect, the charge transition level $\varepsilon(q/q')$ marks the E_F

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

value at which the defect switches from charge state q to q'; below this level one charge state is preferred, and above it the other. The deep or shallow nature of these transition levels determine whether a defect acts as an electron donor, acceptor, or recombination center, and is central to understanding semiconductor doping and device performance. Defect levels are experimentally measured using cathodoluminescence, photoluminescence, optical spectroscopy, or deep-level transient spectroscopy (DLTS).[32] These methods face significant challenges in sample preparation and assigning measured levels to specific defects.[33,34] To overcome this, density functional theory (DFT) is widely used to calculate defect formation energy (E^f) as a function of E_F, defect charge state (q), and chemical potential (μ).[17,35–39] DFT enables identification of donor- and acceptor-type defects, shallow or deep defect levels, type of equilibrium conductivity, defect concentrations, and carrier capture rates.[21,40–43] When an appropriate level of theory is applied, DFT-computed charge transition levels compare well with experiments.[34,40,44,45] However, DFT is computationally expensive and scales poorly with system size, making it difficult to explore the vast configurational space of vacancies, interstitials, antisites, and defect complexes across many compounds and charge states.[46,47]

The prediction of defect properties can be accelerated by integrating DFT simulations with machine learning (ML) approaches such as crystal graph neural networks (GNNs).[48–50] GNNs effectively represent and predict the energies and properties of molecules, polymers, and crystalline materials[51–54] by transforming atomic structures into graphs where atoms are nodes and bonds are edges.[55–57] They learn intricate structural representations to predict properties such as formation or decomposition energy, band gap, and defect formation energy, while reducing the computational cost. In prior work, we used GNNs to predict and screen native defects and functional impurities in group IV, III–V, and II–VI zinc blende semiconductors,[58,59] covering vacancies, interstitials, anti-site, and extrinsic defects. While the models predicted charge-dependent defect formation energies for different chemical potential conditions, several limitations were observed: (1) training on a broad chemical space (34 compounds) led to large errors for specific compositions; (2) models showed good performance on binaries but lower accuracy for alloy systems such as $CdSe_xTe_{1-x}$ and $Cd_xZn_{1-x}Te$;[2,25,27,41,60–62] (3) reliance on modest 64-atom $2 \times 2 \times 2$ supercells limited defect complex modeling; (4) use of the semi-local GGA-PBE functional (GGA: Generalized Gradient Approximation, PBE: Perdew–Burke–Ernzerhof) inherently limited prediction fidelity;[63–65] and (5) dependence on gradient-free optimization with GNN models prevented more efficient gradient-based geometry optimization.

To overcome prior limitations from casting too wide a chemical space, using smaller supercells, and relying on semi-local functionals, we developed a more comprehensive, multi-fidelity methodology. Our approach begins with an initial dataset of bulk and defect configurations spanning Cd/Zn–Te/Se/S binary and multi-nary compounds. These structures were first computed using the PBE functional, providing a baseline set of bulk and defect configurations and their energies, with a substantial portion of the PBE dataset compiled from our previously published works.[40,66] Initial GNN models trained on the PBE dataset[67] served as the foundation for predicting defect properties over a pre-defined defect chemical space containing thousands of vacancies, interstitials, antisites, substitutional defects, and defect complexes. Defect enumeration included all relevant native defects as well as group-V dopants (N, P, As, Sb, Bi) which are promising for achieving p-type conductivity and unintentional impurities such as Cl and O which are known to strongly influence the performance of CdTe and $CdSe_xTe_{1-x}$ solar cells.[68,69] The Cd/Zn–Te/Se/S chemical space was chosen due to its relevance to Se grading, Cd–Zn interfaces, and absorber composition tuning in CdTe solar cells, where exploring all low-energy native and extrinsic defects across these compositions provides a comprehensive dataset for experimental comparison. Although ZnS and ZnSe are not primary absorbers, they remain chemically informative, while CdS functions as an important buffer layer.[5,6,16]

To improve prediction accuracy, active learning was employed to iteratively generate new DFT data and refine the GNN models.[70–75] Following convergence of the active learning scheme, we performed higher-fidelity HSE06 (Heyd–Scuseria–Ernzerhof)[76] calculations on a curated subset of representative PBE-relaxed structures to obtain more accurate band gaps, charge transition levels, and defect formation energies. GNN models, specifically using the M3GNet[55] architecture for machine learning force fields (MLFFs), were then trained on the HSE06 data and subsequently used for new predictions. Our complete methodology is summarized as: PBE data collection → initial PBE GNN model training → active-learning–driven expansion of the PBE dataset → HSE06 refinement of a subset of the PBE dataset → training MLFF models at HSE06 accuracy. The next few sections describe our methodology and results in detail, highlighting the following major contributions of this work:

- Construction of the largest unified HSE06 defect dataset across Cd/Zn–Te/Se/S compositions, including native and extrinsic defects, and defect complexes, simulated in five charge states.
- Development of an HSE06 MLFF-based defect geometry optimization workflow that is orders of magnitude faster than full DFT.
- Release of DeFecT-FF, an online nanoHUB tool[77] with the following workflow: input bulk structure + list of defect candidates → generation of defect structures with symmetry breaking → MLFF optimization across five charge states → selection of the lowest-energy configurations → final HSE06+SOC (spin–orbit coupling) calculation.

2 Description of the DFT datasets

The entire Cd/Zn–Te/Se/S defect chemical space is extraordinarily large, as summarized in the SI, Table S1. Treating every symmetry-inequivalent site in a $3 \times 3 \times 3$ (216-atom) cubic zinc blende supercell results in thousands of possible defect

configurations per composition, driven by the combinatorial explosion of native vacancies, self-interstitials, anti-site substitutions, eight types of extrinsic defects (dopants—Cu, As, P, N, Sb, Bi; impurities—Cl, O), and all unordered defect complexes, as illustrated in Fig. S1. Even when restricted to only neutral charge states, the number of required DFT calculations grows to nearly ≈0.9 million for the entire chemical space. Thankfully, the defect structures collected from published literature and our prior work[40,58,66,78] already represent a physically meaningful and chemically rich subset of this much larger defect chemical space. These literature-curated PBE-optimized structures span multiple compositions, charge states, and defect chemistries, providing a robust foundation for training the initial GNN models. We then employed an active-learning workflow to selectively launch new PBE calculations in regions of the chemical space where the ML model exhibited high uncertainty or sparse representation. Details of this dataset-building strategy are provided in the SI.

A carefully selected subset of PBE-optimized structures was further used to perform hybrid HSE06 calculations after adjusting the defect supercell lattice parameters to values from HSE06 optimization, before volume-fixed relaxation, as summarized in the Table SII. The resulting HSE06 dataset captures more than half of the structural and chemical diversity present in the full PBE dataset, and the statistical distribution of this HSE subset is illustrated in Fig. 1. Importantly, all HSE calculations preserved every ionic-relaxation snapshot—not only the final minimum-energy configuration—yielding thousands of intermediate structures with their corresponding energies, forces, and stresses. This provides a significantly enriched dataset that captures full relaxation pathways rather than ground-state endpoints alone. For a complete statistical overview of the PBE and

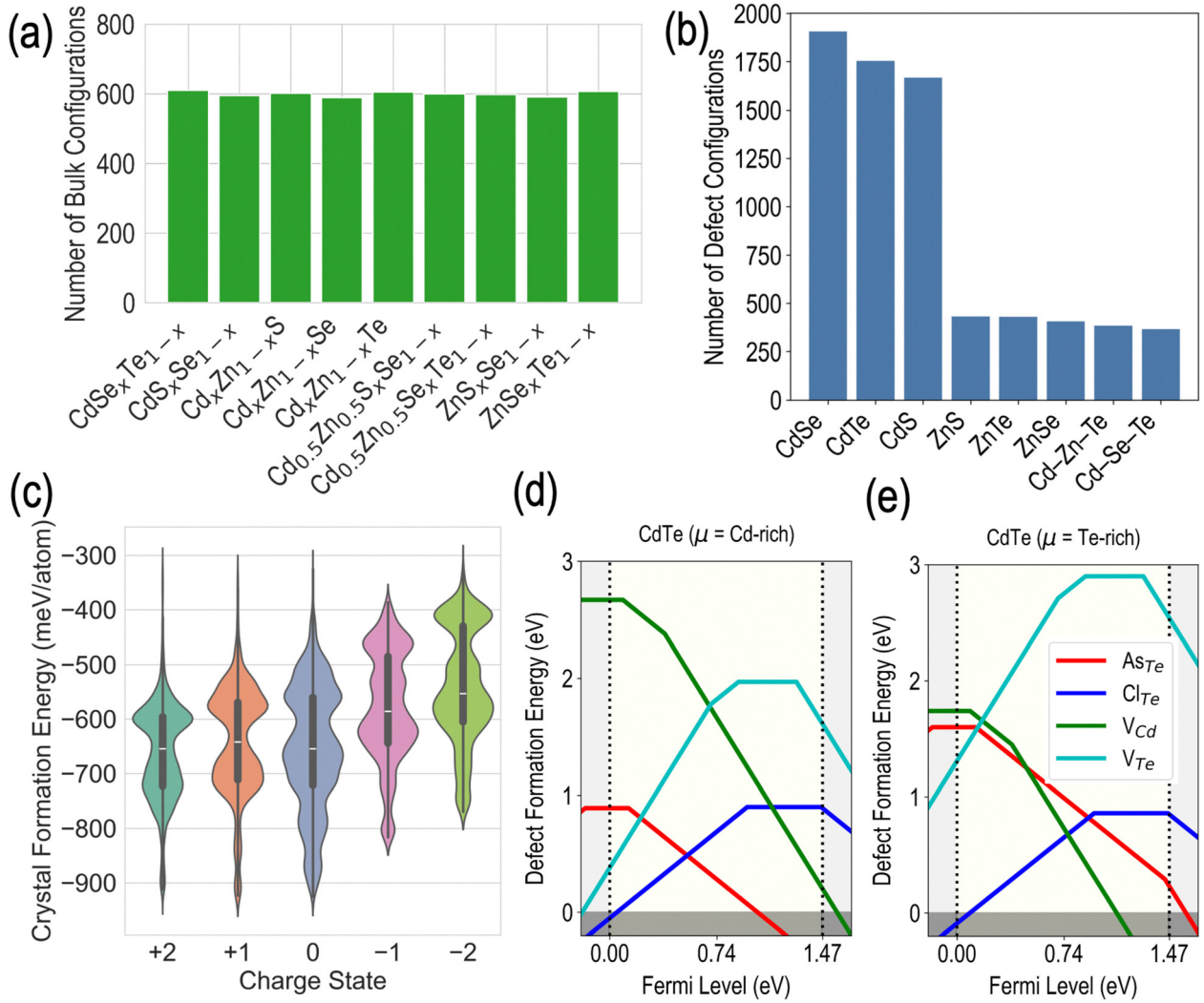

Fig. 1 Statistics of the HSE06 dataset: (a) number of bulk configurations from the $CdSe_xTe_{1-x}$, CdS_xSe_{1-x}, $Cd_xZn_{1-x}S$, $Cd_xZn_{1-x}Se$, $Cd_xZn_{1-x}Te$, $Cd_{0.5}Zn_{0.5}S_xSe_{1-x}$, $Cd_{0.5}Zn_{0.5}Se_xTe_{1-x}$, ZnS_xSe_{1-x}, and $ZnSe_xTe_{1-x}$ compositions. (b) Distribution of defect configurations across the Cd–chalcogen and Zn–chalcogen binaries and ternaries. (c) Violin plots of crystal formation energies (meV per atom) for the entire dataset across five charge states (+2 to −2). (d) and (e) Defect formation energy diagrams for CdTe under Cd-rich and Te-rich conditions from HSE06 functional, highlighting the relative stability of key native (V_{Cd}, V_{Te}) and extrinsic defects (As_{Te}, Cl_{Te}).

HSE datasets across all charge states, readers are referred to Tables SIII–SV. Additional dataset visualization is provided in Fig. S2–S4. In summary, our data generation pipeline proceeds as follows: literature data collection → initial PBE GNN model training → active-learning–driven expansion of the PBE dataset → HSE06 refinement of a subset of the PBE dataset.

DFT calculations were performed using VASP[79] on the Negishi cluster at Purdue University, utilizing nodes equipped with two AMD EPYC 7763 ''Milan'' CPUs @ 2.2 GHz (128 cores per node, 256 GB memory). HSE06 relaxation jobs were run using 512 cores (4 nodes). The MLFF relaxations were performed on a single compute node with 16 cores. Under these conditions, a single HSE06 geometry optimization for a charged defect in a 216-atom supercell requires approximately 4096 core-hours, whereas the DeFecT-FF relaxation (models described in the next section) completes in approximately 0.5 core-hours, representing a speedup exceeding four orders of magnitude.

3 MLFF models trained at hybrid functional accuracy

We initially trained models on the PBE dataset using the ALIGNN framework[80,81] and then employed an active learning strategy to launch new DFT calculations by targeting regions of the chemical space with largest prediction uncertainty. We then refined a representative subset of the PBE-optimized structures using the HSE06 functional and used these data to train an M3GNet-based[55] MLFF, using DFT-derived configurations, energies, forces, and stresses. Readers are referred to the SI for additional details on the ALIGNN training procedure and the active learning workflow used in this work. These discussions are supplemented by a series of figures in the SI (Fig. S5–S13), which collectively illustrate the active learning pipeline, model transferability, ALIGNN-based structural optimization, comparisons with other MLFF models, and MLFF models trained on the PBE dataset. The active learning workflow[82] employs an ensemble of ALIGNN models,[80] with the standard deviation of their energy predictions serving as the uncertainty metric. In each active learning batch, the top 200 structures with the highest prediction uncertainty (generally 5–10 meV per atom standard deviation) are selected for additional DFT calculations. Convergence is assessed by monitoring the fraction of newly queried structures falling within the model's confidence interval and the saturation of test set RMSE. After 3–4 iterations, both metrics indicated convergence. Additional details on the active learning pipeline, including sensitivity analysis, are provided in the SI (Fig. S5–S13).

HSE06 calculations were performed using Γ-point only,[83] with a reduced plane-wave energy cutoff of 400 eV. The convergence thresholds for geometry optimization were set to 10^{-6} eV for energy and 0.01 eV Å^{-1} for forces. Fig. 1 shows the statistics of the compiled HSE06 dataset in terms of the number of bulk Cd/Zn–Te/Se/S composition structures, and different types of defects. The HSE06 dataset represents 53.4% of the GGA dataset for bulk (q = 0) structures, and 63.8%, 71.6%, 79.5%, 72.4%, and 65.8% for defect structures in charge states q = +2, q = +1, q = 0, q = −1, and q = −2, respectively. Despite the smaller size, it remains well representative of the defect types and structural diversity of the entire chemical space. Violin plots showing the spread of the crystal formation energy (CFE, a per-atom energy difference between the crystal and its constituent atoms) values in the HSE06 dataset are presented in Fig. S4. CFE is defined as the per-atom energy difference between the total energy of the crystal and the sum of elemental reference energies of its constituent elements, *i.e.*, $\mathrm{CFE} = \left(E_{\mathrm{crystal}} - \sum_i n_i E_{\mathrm{ref,i}}\right) \Big/ N$, where $E_{\mathrm{ref,i}}$ is the reference energy of element i and N is the total number of atoms in the supercell. The reference states for all elements are taken as the lowest-energy phases from the Materials Project.[84]

Several MLFF frameworks have been developed for materials property prediction, including pretrained universal models such as MACE,[85] CHGNet,[86] and M3GNet.[55] However, these models are trained predominantly on neutral bulk structures from databases such as the Materials Project,[84] and consequently generalize poorly to charged defect configurations in semiconductors. DeFecT-FF addresses this gap through four key innovations: (i) charge-state-resolved models that explicitly capture the structural and energetic signatures of defects in five charge states; (ii) a multi-fidelity active learning pipeline that efficiently bridges PBE and HSE06 levels of theory; (iii) training on defect-specific data including intermediate relaxation snapshots, symmetry-broken geometries from ShakeNBreak,[45] and defect complexes; and (iv) deployment as an end-to-end nanoHUB tool for community use.

Before training the MLFF models, we first evaluated the performance of state-of-the-art pretrained force-field frameworks, namely MACE,[85] CHGNet[86] and M3GNet,[55] by applying them directly to our PBE dataset. Although these models have demonstrated strong predictive capabilities on their training domains, they generalized poorly to the chemically diverse Cd/Zn–Te/Se/S systems investigated in this work as summarized in Table SVI. Root mean square error (RMSE) in CFE prediction ranged from 60 to 100 meV per atom, which it will turn out are much larger than errors from fine-tuned models. These shortcomings indicate that the pretrained models lack sufficient exposure to the chalcogenide defect chemical space considered here. Consequently, this motivated the development of a dataset-specific M3GNet-based MLFF trained on DFT-derived configurations, energies, forces, and stresses, enabling the level of accuracy required for reliable modeling of defect thermodynamics and structural relaxations.

Parity plots for M3GNet-MLFF models trained on the HSE06 dataset are pictured in Fig. 2(a)–(c), respectively for charge states q = +1, q = 0, and q = −1. Models for the q = +2 and q = −2 charge states are presented in Fig. S15. Each parity plot compares the HSE06-computed CFE with MLFF-predicted values across different categories: bulk (pristine supercells without defects), and defects (bulk supercells containing a single defect or defect complex). Despite the reduced dataset

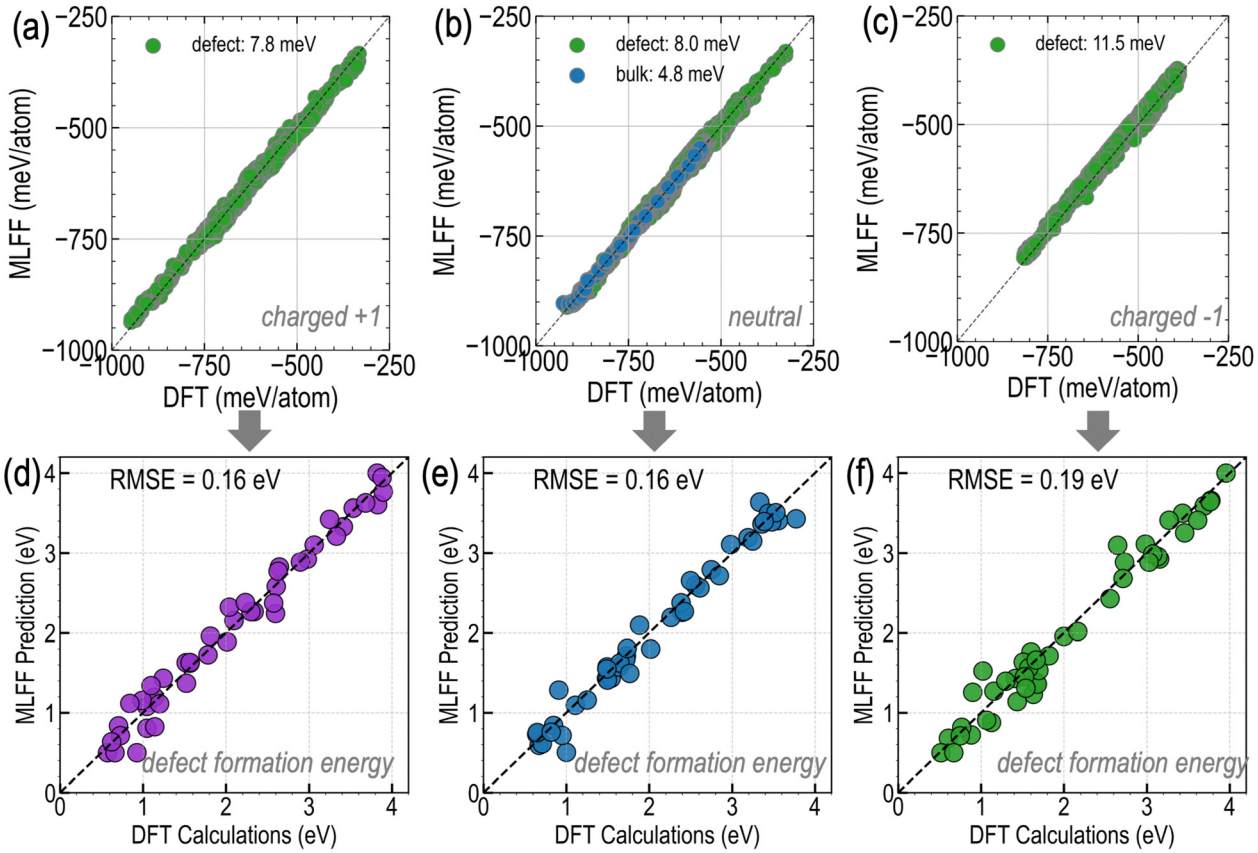

Fig. 2 (a)–(c) Parity plots comparing crystal formation energies from DFT and MLFF predictions, for three representative charge states: (a) q = +1, (b) q = 0 (neutral), and (c) q = −1. The MLFF accurately reproduces the DFT energies with small errors, as indicated by the RMSE values shown in each panel. (d)–(f) Defect formation energies under Cd-rich condition computed using MLFF predictions for a subset of the defect configurations shown in panels (a)–(c), compared against values from full DFT. The MLFF defect energies were obtained by adding DFT reference energies and applying charge corrections to the MLFF-predicted total energies.

size compared to the GGA dataset, the HSE-MLFF models achieve very good accuracy with low RMSE values across different structure types. The q = 0 test set prediction RMSE ranges from 4.8 meV per atom for bulk structures to 7.8 meV per atom for defect structures. These errors are similar for q = +1 and q = −1 defect structures and remain below 12 meV per atom for all cases, which is quite reasonable given the range of CFE values in the dataset. We also simulated a limited number of CdTe dislocation core structures[87,88] and CdTe/ZnTe interface configurations with selected defects and included them in the training dataset. The model performance for these structures is shown in Fig. S14 and S15. These results are not discussed in detail in the main text because the number of data points corresponding to dislocation cores and interfaces is relatively small.

The training dataset spans the entire Cd/Zn–Te/Se/S chemical space with all native defect types and extrinsic defect species across five charge states, covering 14 individual compounds that include 6 binaries and 8 ternary alloys. Table SVIII provides a breakdown of model accuracy by defect type, showing consistent performance across vacancies (9.5 meV per atom), extrinsic substitutions (8.6 meV per atom), anti-site substitutions (8.1 meV per atom), and interstitials (7.6 meV per atom). Active learning specifically targeted underrepresented regions to ensure broad coverage of the defect chemical space.

Fig. S16 shows the MLFF predictions for neutral defect structures separately for single defects and defect complexes, revealing low RMSE values of 8.14 meV per atom and 9.23 meV per atom respectively. This suggests that the MLFF effectively captures both localized and collective defect relaxations, even for configurations with multiple defects. Even though MLFF prediction of crystal formation energy is highly accurate for all bulk and defect structures, a true evaluation of the prediction for defect configurations involves comparing defect formation energy (and defect transition level) predictions from MLFF and HSE06. We used the MLFF models to optimize a selected set of defects structures. For each defect, the MLFF was used to compute the total energy entering the standard defect formation energy expression:

$$\Delta E_f(D^q) = E_{tot}(D^q) - E_{tot}(\text{bulk}) + \sum_i n_i\mu_i + q(E_F + E_{VBM}) + E_{corr}$$

Here, $E_{tot}(D^q)$ and E_{tot}(bulk) are the total energies of the defect and pristine supercells respectively, n_i and μ_i denote the stoichiometric changes and elemental chemical potentials, q is the charge state, E_{VBM} is the valence band maximum energy,

E_F is the Fermi level through the band gap, and E_{corr} is the correction energy[17] which accounts for spurious electrostatic interactions arising from the periodic repetition of charged defects and the compensating background charge in finite supercells. In this work, E_{corr} is evaluated using the Freysoldt[17] charge correction scheme, which separates long-range Coulomb interactions from short-range defect-induced potentials and aligns the electrostatic potential between defect and bulk calculations.[17] Additional methodological details and convergence tests for the charge correction are provided in the SI. Importantly, the reference energy μ_i, E_{VBM} and E_{corr} were evaluated from DFT and added directly to the MLFF-derived bulk and defect energies.

Fig. 2(d)–(f) present parity plots for defect formation energy corresponding to $q = +1$ (at $E_F = 0$), $q = 0$, and $q = -1$ (at $E_F = 0$). Across the entire validation set, the RMSE in defect formation energies obtained using MLFF-optimized geometries remains below 0.20 eV, demonstrating that the MLFFs are sufficiently accurate in capturing both structural and energetic trends for charged and neutral defects. To further assess the influence of charge corrections on MLFF-derived defect energetics, we compared three approaches: (i) using MLFF-predicted defect formation energies without any charge correction; (ii) using MLFF defect formation energies corrected using a simple average offset of 0.20 eV for $q = +2$, 0.10 eV for $q = +1$, 0.10 eV for $q = -1$, and 0.20 eV for $q = +2$ defects; and (iii) adding known charge correction energies from DFT to MLFF-predicted defect formation energies (Fig. 2(d)–(f)).

The average charge-correction offsets ($\sim$0.20 eV for $|q| = 2$, $\sim$0.10 eV for $|q| = 1$) were derived empirically from extensive defect calculations across the Cd/Zn–Te/Se/S chemical space. Their near-uniformity arises from the similar supercell geometries and the narrow range of dielectric constants ($\sim$7–10) in this class of materials. We note that these offsets are specific to the present chemical space and may not be applicable to other chemistries with different dielectric properties, crystal structures, or supercell sizes. For applications beyond the Cd/Zn chalcogenide systems, explicit DFT-based Freysoldt corrections[17,89] are necessary.

The comparison reveals that while uncorrected MLFF prediction shows errors close to 0.3 eV for all charge states, applying the average offset brings this error down closer to 0.2 eV which is similar to the error from adding known correction values, as listed in Table SVII. Parity plots comparing MLFF defect formation energy with DFT values for all three approaches are shown in Fig. S17. Fig. S18 compares defect charge transition levels predicted by the MLFF with DFT reference values. Applying an average charge correction value leads to close agreement between MLFF- and DFT-predicted defect transition levels, with RMSE values of 0.25 eV, 0.23 eV, 0.22 eV, and 0.27 eV for the $\varepsilon(+2/+1)$, $\varepsilon(+1/0)$, $\varepsilon(0/-1)$, and $\varepsilon(-1/-2)$ transitions, respectively, whereas the errors when applying the DFT-based charge correction values are 0.22 eV, 0.21 eV, 0.20 eV, and 0.24 eV. The increase in RMSE from meV per atom (for CFE) to $\sim$0.2 eV (for defect formation energies) arises because the defect formation energy expression combines MLFF-predicted supercell energies with independently computed DFT-derived quantities (chemical potentials, E_{VBM}, charge corrections). The MLFF prediction errors for the defect and bulk supercells do not cancel perfectly due to distinct local atomic environments around the defect site. The residual $\sim$0.2 eV error reflects this imperfect cancellation combined with the reference-frame mismatch between MLFF and DFT contributions to the defect formation energy expression. This defect formation energy error is still reasonable and enormously useful for quick prediction and screening.

To assess the reliability of our MLFF models relative to full DFT, we randomly selected 100 representative bulk and defect configurations and relaxed each structure using both methods. The DFT- and MLFF-optimized geometries were then compared using SOAP descriptors,[90,91] which provide a rotationally invariant fingerprint of the atomic environments. These high-dimensional descriptors were projected onto a two-dimensional PCA[92] (principal component analysis) space, allowing direct visualization of structural similarity. In Fig. S19, each DFT structure (blue) is paired with its corresponding MLFF structure (orange), with a connecting line indicating the degree of agreement. The consistently short line segments demonstrate that the MLFF reproduces DFT relaxation behavior with high accuracy. Some representative examples of MLFF-based defect structure optimizations are shown in Fig. S11(d)–(f) and S14(d)–(f) of the SI.

The MLFF achieves comparable accuracy across the different material families in our chemical space, as summarized in Table SIX. Binary compounds exhibit slightly lower RMSE (7–8 meV per atom) compared to ternary alloys (8–10 meV per atom), reflecting the additional structural complexity from compositional disorder. Since the charge correction (which depends on the dielectric constant) is computed from DFT and applied separately, variations in dielectric properties across materials do not impact the MLFF's structural prediction accuracy. Ultimately, there are a sufficient number of bulk and defect configurations from different compositions in the training dataset to ensure accurate predictions across the entire space.

We note that separate MLFF models are trained for each charge state (from $q = +2$ to -2), with every model learning the configuration-to-energy mapping from the DFT geometry relaxation performed at that specific charge state. Electrostatic effects are thus implicitly encoded in the training data. Long-range periodic image interactions are corrected using the Freysoldt scheme[17] with DFT-derived quantities. The MLFF captures geometry-driven relaxations through local atomic descriptors, while the electronic structure (band edges, charge corrections, SOC effects) is always determined from the final HSE06+SOC single-point calculation. For defects with delocalized charge densities near band edges, the MLFF still provides accurate structural relaxation; however, it should be noted that the electronic characterization of such shallow defects relies entirely on the DFT step rather than the MLFF prediction. MLFF-driven geometry optimization at different charge states significantly reduces the DFT expense, but the subsequent

high-fidelity calculation is necessary to obtain the final defect formation energies.

The total computational investment for developing DeFecT-FF includes DFT data generation and MLFF training. The HSE06 dataset generation required approximately 20 million core-hours across all compositions and charge states. Training each charge-state-specific M3GNet[55] model required approximately 8–12 GPU-hours on a single NVIDIA A100 GPU. Once trained, the MLFF enables rapid predictions at negligible cost (~0.5 core-hours per defect optimization). The upfront investment in data generation and training is amortized over the large number of subsequent predictions: screening hundreds of defect configurations across multiple compositions requires only days of MLFF computation compared to years of equivalent DFT effort. For users wishing to apply DeFecT-FF to new but related chemistries, fine-tuning the pretrained models on a modest set (~50–100 structures) of new DFT calculations is expected to be sufficient, requiring only a few GPU-hours of additional training.

4 New predictions with the MLFF models: case studies of important defects

The MLFF models can now be used to optimize any given defect configuration in different charge states with near–hybrid-functional accuracy. Following MLFF optimization, final HSE06+SOC single-point calculations must be performed to obtain reliable defect formation energy diagrams. Fig. 3 illustrates the overall workflow of the DeFecT-FF framework[77] in determining the lowest energy symmetry-broken defect configuration followed by a high-fidelity understanding of the defect thermodynamics. Fig. S20 shows the workflow of the DeFecT-FF[77] web tool we created on the nanoHUB platform to enable efficient creation of defect structures, MLFF optimization, and visualization of defect formation energy diagrams. In the next subsections, we present a few case studies demonstrating the application of this workflow to determine the relative stability and charge transition levels of important defects in selected Cd/Zn–Te/Se/S compositions which were not entirely part of the MLFF training dataset.

4.1 As + Cl defect complex in $CdSe_{0.12}Te_{0.88}$

As is a commonly used p-type dopant in Se-alloyed CdTe and Cl is a common impurity arising from $CdCl_2$ treatment, which makes it important to investigate defect complexes of As and Cl in representative CdSeTe compositions. We applied the DeFecT-FF framework to simulate the As_{Se} + Cl_{Se} substitutional defect complex in the compound $CdSe_{0.12}Te_{0.88}$, across five possible charge states. An example configuration is illustrated in Fig. 4(a); Se alloying starting from a 216-atom CdTe cubic supercell is first accomplished using the special quasirandom structures (SQS) approach,[93] following which As and Cl substitution is incorporated in multiple possible locations to eventually yield the preferred combination. For each charge state between q = +2 and q = −2, ten symmetry-broken initial structures were generated using the ShakeNBreak protocol,[44] enabling exploration of a diverse set of competing local geometries. Each of these structures was relaxed using DeFecT-FF to identify the lowest-energy configuration prior to high-fidelity DFT refinement.

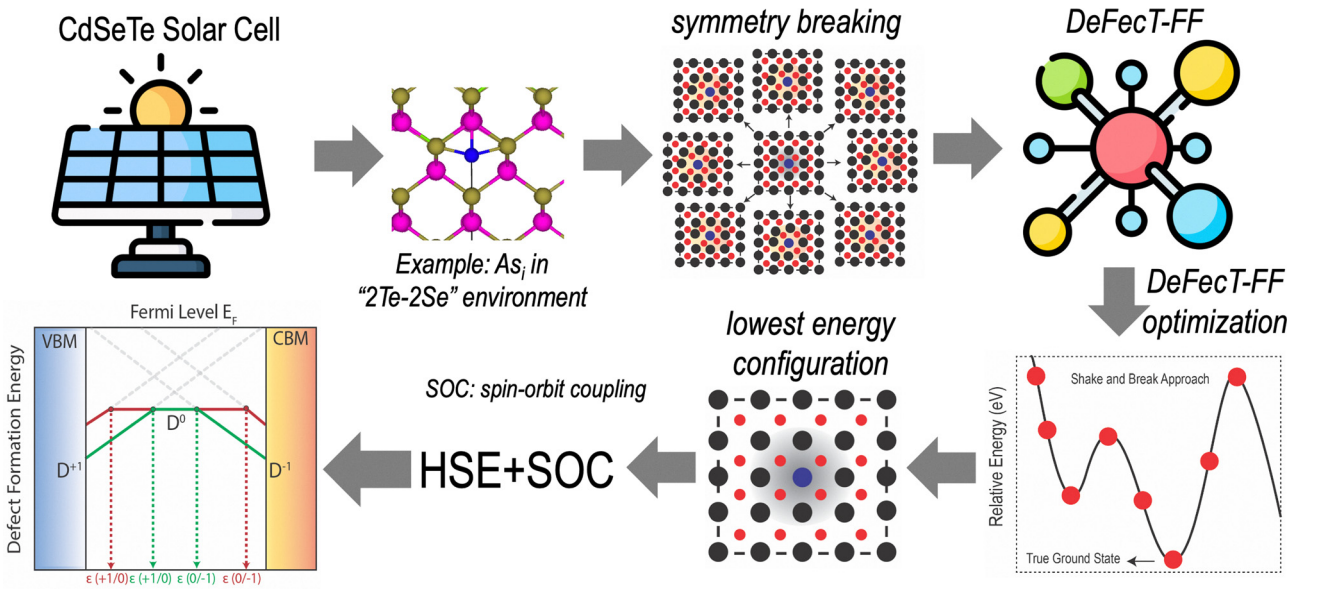

Fig. 3 Workflow for accelerated defect predictions using the DeFecT-FF framework. An initial defect structure (example: As_i in a mixed "2Te–2Se" local environment) is constructed and passed through the ShakeNBreak[44] symmetry-breaking procedure to generate a diverse set of competing defect geometries. These distorted configurations are rapidly relaxed using the rigorously optimized machine-learned force field to identify the lowest-energy structure prior to high-fidelity DFT calculations. The optimized geometry is then used to perform static HSE+SOC calculation, yielding accurate defect formation energy diagrams.

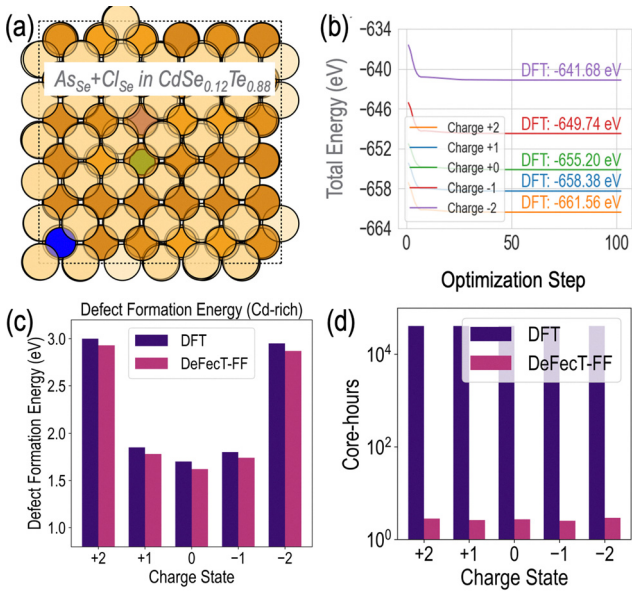

Fig. 4 Benchmarking DeFecT-FF for a selected defect complex in $CdSe_{0.12}Te_{0.88}$. (a) Visualization of the As_{Se} + Cl_{Se} complex in the $CdSe_{0.12}Te_{0.88}$ alloy supercell. (b) Total energy relaxation profiles for different charge states, comparing the converged DFT energies with the DeFecT-FF-relaxed energies. (c) Defect formation energies under Cd-rich conditions for charge states +2 to −2, showing close agreement between DFT and DeFecT-FF predictions. (d) Computational cost, measured in core-hours, highlighting the significant reduction in wall-time achieved when using DeFecT-FF instead of full DFT relaxations.

The total energy relaxation profiles for different charge states are shown in Fig. 4(b). The DeFecT-FF structural optimizations converge smoothly and yield geometries that lie very close to those produced by DFT calculations. After DeFecT-FF relaxation, a single-shot HSE06+SOC calculation is performed on the predicted lowest-energy geometry for each charge state to accurately compute defect formation energies. The resulting defect formation energies (at E_F = 0) under Cd-rich conditions from DFT and DeFecT-FF are compared in Fig. 4(c). Across all charge states, the agreement is excellent, with typical deviations well below 0.1–0.2 eV.

Thus, the DeFecT-FF geometries provide a sufficiently accurate structural foundation for defect thermodynamics for complexes in an alloyed CdSeTe composition. The computational savings are substantial: a single HSE06 relaxation of a charged defect in a 3 × 3 × 3 supercell requires approximately 512 cores multiplied by 8 to 9 hours per configuration, corresponding to a total of nearly 4096 core-hours. In contrast, the DeFecT-FF relaxations require only about (2/60) × 16 core-hours per configuration, which is approximately 0.5 core-hours. The speedup therefore exceeds four orders of magnitude, as presented in Fig. 4(d). We expect these trends to hold for all types of defect complexes in alloyed 3× 3 × 3 supercells and there is confidence in DeFecT-FF reaching close agreement with hybrid DFT at a fraction of the cost.

4.2 As and Cl defects across $CdSe_xTe_{1-x}$ compositions

Next, we simulated multiple substitutional defects of As and Cl (including complexes) in 3 × 3 × 3 supercells of a series of $CdSe_xTe_{1-x}$ compositions (x = 0, 0.06, 0.12, 0.25). Using the Doped[94] package, we introduced defects As_{Te}, As_{Se}, Cl_{Te}, Cl_{Se} and the As_X + Cl_X double defect complexes (where X denotes the preferred anion site, Te or Se). Symmetry-breaking operations were then applied *via* the ShakeNBreak protocol, enabling the sampling of a diverse set of competing configurations (Fig. S21). Hundreds of structures for these substitutional defects across the $CdSe_xTe_{1-x}$ compounds were relaxed with the DeFecT-FF models for different charge states until the maximum force fell below $<10^{-2}$ eV Å^{-1}. Finally, single-shot HSE06+SOC calculations were performed to obtain accurate defect formation energy.

The band gaps computed using HSE06+SOC (with a modified mixing parameter of α = 0.31) for CdTe, $CdSe_{0.06}Te_{0.94}$, $CdSe_{0.12}Te_{0.88}$, and $CdSe_{0.25}Te_{0.75}$ are respectively 1.5 eV, 1.41 eV, 1.38 eV, and 1.30 eV; these values are used to place the E_F bounds for the defect formation energy diagrams. The VBM for each composition was obtained from the bulk calculation at the HSE+SOC level using a 2 × 2 × 2 k-mesh for the corresponding 216-atom 3 × 3 × 3 supercell. The charge-dependent defect formation energies additionally yield the charge transition levels as described below:

$$\varepsilon(q/q') = \frac{\Delta E_f(D^q; E_F = 0) - \Delta E_f(D^{q'}; E_F = 0)}{q' - q}$$

This transition level marks the E_F position at which charge states q and q' are in equilibrium. Fig. 5 presents selected defect formation energy diagrams and the relevant transition levels for As_X, Cl_X, and As_X + Cl_X defects across the $CdSe_xTe_{1-x}$ series, with E_{VBM} set to 0 eV; X represents either Te or Se. Incorporation of Se is observed to deepen the As_X 0/−1 acceptor level despite the band gap going down from CdTe to $CdSe_{0.25}Te_{0.75}$, in agreement with recent experimental studies.[95] The Cl_X + 1/0 donor level remains deep in the band gap in all cases, around 1 eV from the VBM, while the As_X + Cl_X defect complex, interestingly, creates a 0/−1 acceptor level closer to the conduction band edge which becomes shallower with more Se content due to the lowering of the CBM. The defect energy diagrams in Fig. 5(a) and (b) show the prevalence of the neutral state for the defect complex in the band gap, while As_X and Cl_X respectively create low energy acceptor and donor defects which pin the equilibrium E_F (obtained by applying charge-neutrality conditions) around the middle of the band gap.

The case studies in this section serve as out-of-distribution validation tests:[96] the defect configurations examined (As + Cl complex in $CdSe_{0.12}Te_{0.88}$) were not included in the MLFF training dataset. To further quantify OOD performance, we evaluated the model on two compositions entirely absent from the training set: $CdSe_{0.12}Te_{0.88}$ and $CdSe_{0.06}Te_{0.94}$. As shown in Table SX, the RMSE values (12–13 meV per atom) are moderately higher than in-distribution errors but confirm reasonable generalization. We note that systematic uncertainty quantification for the deployed MLFF models (*e.g.*, *via* ensemble predictions or Monte Carlo dropout) remains a direction for future development.

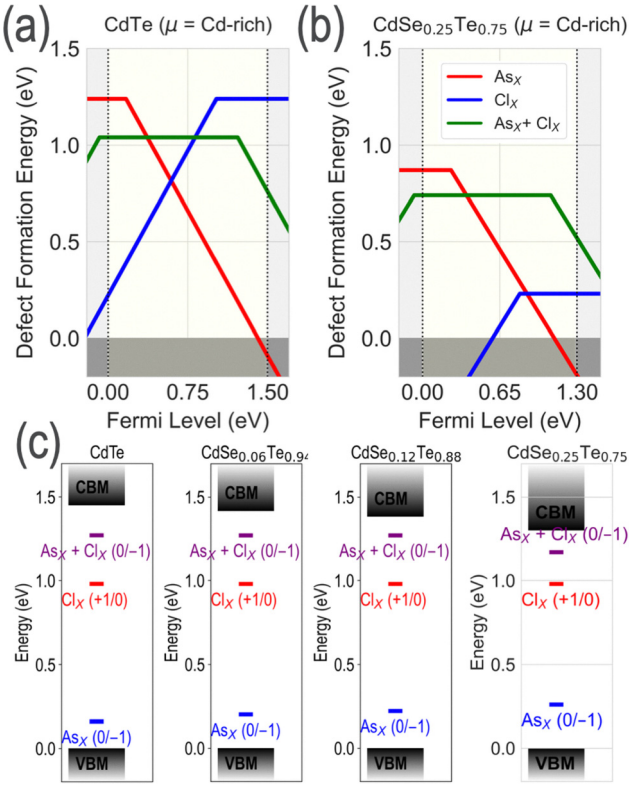

Fig. 5 Defect formation energy diagrams for As_X, Cl_X, and $As_X + Cl_X$ defects in (a) CdTe and (b) $CdSe_{0.25}Te_{0.75}$, under Cd-rich conditions; X = Te or Se. (c) Defect charge transition levels for As_X, Cl_X, and $As_X + Cl_X$, computed for different $CdSe_xTe_{1-x}$ compositions (x = 0.0, 0.06, 0.12, 0.25). Blue lines indicate the As_X (0/−1) acceptor level, red lines show the Cl_X (+1/0) donor level, and purple lines show the $As_X + Cl_X$ (0/−1) level. For each compound, the VBM is placed at E_F = 0 eV and the CBM (conduction band minimum) is placed at the value of the computed band gap. All results are from HSE06+SOC calculations performed after DeFecT-FF optimization.

4.3 Native defects and nitrogen impurities in ZnTe

Motivated by experimental evidence from X-ray photoelectron spectroscopy (XPS)[97–102] indicating N incorporation in ZnTe,[7,103–108] we employed the DeFecT-FF workflow to systematically investigate both native point defects and N-related defects in ZnTe. A 3 × 3 × 3 ZnTe supercell (cubic zinc blende phase) was first fully relaxed using the HSE06 functional prior to defect introduction; its band gap was computed to be 2.2 eV from HSE06+SOC. The defect set included vacancies, interstitials, and antisite defects (V_{Zn}, V_{Te}, Zn_i, Te_i, Zn_{Te}, Te_{Zn}), as well as the following N defects: N_i, N_{Te}, $N_i + N_i$, and $N_{Te} + N_i$. To ensure thorough exploration of the potential energy landscape, we applied ShakeNBreak[44,45] to induce perturbations, enabling the sampling of a diverse set of competing configurations. Among the hundreds of N-related configurations evaluated, the $N_i + N_i$ defect complex emerged as the most energetically favorable, a finding further validated through additional HSE06+SOC calculations. Fig. 6(a) illustrates the DeFecT-FF structural optimization and energy convergence for $N_i + N_i$ in ZnTe, and Fig. 6(b) presents the HSE06+SOC computed

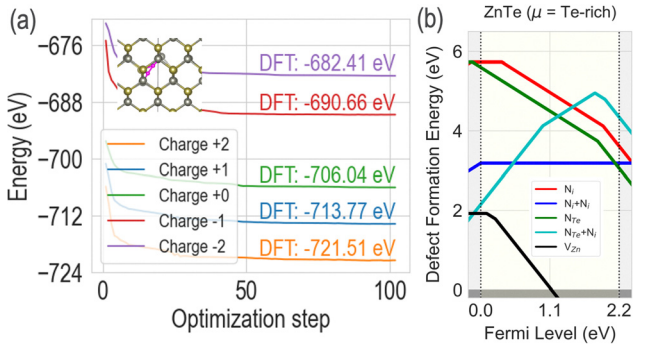

Fig. 6 (a) Energy as a function of optimization steps during DeFecT-FF relaxation of a ZnTe supercell with the double N interstitial ($N_i + N_i$) defect complex. The inset shows the relaxed configuration with N atoms in red. (b) Defect formation energies in ZnTe under Te-rich conditions computed using HSE06+SOC on top of the HSE-MLFF optimized configurations.

defect formation energy diagram for different N-related defects in ZnTe.

To clarify the role of the MLFF in the defect formation energy workflow: the primary computational bottleneck is the geometry optimization of defect supercells, which requires many ionic relaxation steps using the HSE06 functional (roughly 4000 core-hours per defect). DeFecT-FF replaces this step with a fast MLFF-based relaxation (~0.5 core-hours), after which a single-point HSE06+SOC calculation is performed on the optimized geometry. The band gap and valence band edge are obtained from DFT on the bulk semiconductor and only need to be computed once per composition. These values, along with available chemical potentials, are combined with MLFF-predicted energies and assumed charge correction energies to obtain MLFF-based defect formation energies. Thus, the MLFF eliminates the need for iterative and expensive DFT defect geometry optimization while enabling rapid screening of many competing configurations, reducing the total cost by over four orders of magnitude.

4.4 Simulating defects at finite temperature

Defect formation energies in semiconductors are typically evaluated at T = 0 K using DFT, neglecting vibrational entropy effects that can become important under realistic growth and operating conditions. While a full finite-temperature free-energy treatment is beyond the scope of this work, we demonstrate that DeFecT-FF enables finite-temperature molecular dynamics (MD)[109,110] simulations that provide a physically grounded pathway toward incorporating such effects in future studies. As a representative example, we consider the As_{Te} defect in CdTe. The defect structure was first relaxed using DeFecT-FF, followed by finite-temperature microcanonical (NVE) ensemble MD simulations for the neutral charge state (q = 0) at T = 300 K using a 1 fs time step and a total simulation length of 100 000 steps.[111–113] Fig. 7(a) shows the energy as a function of simulation time, demonstrating excellent energy conservation over the 100 ps trajectory and confirming the numerical stability of DeFecT-FF.

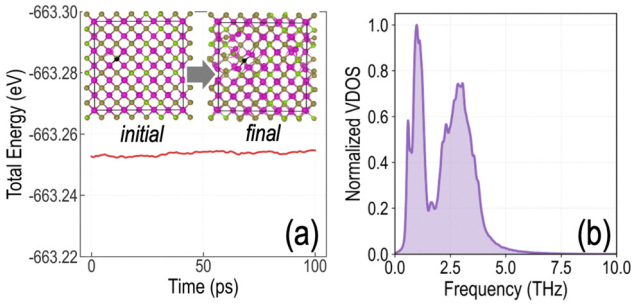

Fig. 7 (a) Total energy as a function of simulation time for a microcanonical (NVE) ensemble molecular dynamics trajectory of the As_{Te} defect in CdTe, demonstrating excellent energy conservation over a 100 ps run. The inset shows the atomic structure before (initial) and after (final) MD simulation, with the defect site highlighted. (b) Normalized vibrational density of states (VDOS) obtained from the velocity autocorrelation function of the same trajectory, revealing dominant low-frequency vibrational modes below 4 THz that are characteristic of defect-localized and heavy-atom vibrations.

Beyond validating stability, finite-temperature MD offers key physical advantages over static relaxations by enabling symmetry-breaking driven by thermal fluctuations and realistic atomic motion. During the MD simulations, atomic velocities $v_i(t)$ were recorded at each time step, where i labels atoms in the supercell and N is the total number of atoms. From these trajectories, we computed the velocity autocorrelation function as follows:[114]

$$C_{vv}(t) = \frac{1}{3N}\sum_{i=1}^{N}\langle v_i(0)\cdot v_i(t)\rangle, \quad (1)$$

and obtained the vibrational density of states (VDOS), $g(\omega)$, using Fourier transformation:

$$g(\omega) = \int_0^\infty C_{vv}(t)e^{-i\omega t}dt, \quad (2)$$

Here, $\omega = 2\pi f$ is the vibrational angular frequency. Fig. 7(b) shows the normalized VDOS derived from the MD trajectory. The dominant low-frequency modes below approximately 4 THz are characteristic of defect-localized vibrations and reflect vibrational softening induced by the As_{Te} defect and the presence of heavy atoms in CdTe. Beyond confirming numerical stability, the finite-temperature MD simulation provides a few physical insights and opportunities: (i) the VDOS reveals defect-specific vibrational signatures, including low-frequency modes characteristic of defect-localized vibrations; (ii) thermal fluctuations enable dynamical exploration of the local configurational landscape,[17] potentially accessing lower-energy structures through symmetry-breaking pathways inaccessible at 0 K; and (iii) the VDOS provides direct access to vibrational entropy contributions needed for computing temperature-dependent defect formation free energies, establishing a foundation for future investigations of defect thermodynamics under realistic conditions which will be addressed in follow-up contributions from our group.

5 Conclusions

$CdSe_xTe_{1-x}$ solar cells are fundamentally constrained by defect physics: deep-level nonradiative centers from native defects and impurities limit open-circuit voltage, dopants such as Cu and As often lead to unhelpful complexes, and extended defects at interfaces and grain boundaries act as sinks for charge and sites for defect clustering. While hybrid-functional DFT remains the gold standard for resolving these mechanisms, its cost prevents exhaustive exploration across alloy compositions, charge states, and structural motifs. To overcome these barriers, we developed the DeFecT-FF framework, a crystal graph-based active learning-driven MLFF model trained on data from both semi-local GGA and hybrid HSE06 calculations for thousands of charged and neutral structures spanning the Cd/Zn–S/Se/Te chemical space, with a wide variety of native and extrinsic defects and defect complexes considered. DeFecT-FF predicts energies and forces across charge states, enabling rapid geometry optimization and defect formation energy evaluation. We demonstrated the utility of these models by identifying low energy configurations of device-relevant defects and performing HSE06+SOC calculations to understand their energetics and defect levels.

In practice, the DeFecT-FF framework reduces single defect optimization time from at least ~8–9 h (HSE06) to ~1–2 min while retaining near-DFT accuracy, transforming comprehensive, composition- and charge-resolved defect surveys from intractable to routine. The term "near-DFT accuracy" refers specifically to the MLFF achieving RMSE values of 5–10 meV per atom in crystal formation energy and < 0.20 eV in defect formation energy relative to HSE06 DFT. "High-fidelity" refers to the use of the HSE06+SOC for final single-point calculations.

We have deployed this framework as part of a Jupyter notebook-based nanoHUB tool which will allow users to upload CIF files of Cd/Zn–Te/Se/S structures, auto-generate relevant defects or complexes, and compute their defect formation energies as functions of Fermi level and chemical potentials conditions, bypassing expensive first principles workflows. Together, these advances provide a scalable, charge-aware pathway to map defect landscapes in chemistries relevant to CdSeTe solar cell devices and beyond, accelerating the dopant/process optimization and ultimately closing the voltage deficit in this important thin-film photovoltaic platform.

Author contributions

A. M.-K. conceived and planned the research project and procured research funding. DFT computations and MLFF training tasks were performed by M. H. R. and M. B.; M. H. R. took the lead on writing; M. B. and A. M.-K. contributed in editing and shaping the manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

All data generated and analyzed in this work, including atomic structures, defect configurations, total energies, and machine-learning force-field models, are available through the nanoHUB platform at: https://nanohub.org/resources/defectdatabase.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d6cp00170j.

Acknowledgements

This material is 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. A. M. K. additionally acknowledges support from Argonne National Laboratory under sub-contracts 21090590 and 22057223, from DOE EERE. This research used resources from the the Center for Nanoscale Materials (CNM) at Argonne National Laboratory. Work performed at the CNM, a U.S. Department of Energy Office of Science User Facility, was supported by the U.S. DOE, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. This work also utilized the Anvil cluster at Purdue through allocation MAT230030 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants 2138259, 2138286, 2138307, 2137603, and 2138296. The authors would like to acknowledge discussions with Dr Mariana Bertoni at Arizona State University, Dr Yanfa Yan at University of Toledo, Dr Mike Scarpulla at University of Utah, and researchers at the National Renewable Energy Laboratory. We also acknowledge the Rosen Center for Advanced Computing (RCAC) clusters at Purdue University for further computational support.

References

1 S. Rojsatien, A. Mannodi-Kanakkithodi, T. Walker, T. Nietzold, E. Colegrove, B. Lai, Z. Cai, M. Holt, M. K. Chan and M. I. Bertoni, *Radiat. Phys. Chem.*, 2023, **202**, 110548.
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, *Chem. Mater.*, 2023, **35**, 9935–9944.
3 M. H. Rahman and A. Mannodi-Kanakkithodi, *Comput. Mater. Sci.*, 2025, **249**, 113654.
4 M. H. Rahman, I. Agrawal and A. Mannodi-Kanakkithodi, *2025 IEEE 53rd Photovoltaic Specialists Conference (PVSC)*, 2025, pp. 0717–0719.
5 M. Gloeckler, I. Sankin and Z. Zhao, *IEEE J. Photovolt.*, 2013, **3**, 1389–1393.
6 T. Nideep, M. Ramya and M. Kailasnath, *Superlattices Microstruct.*, 2020, **141**, 106477.
7 E. Menéndez-Proupin, M. Casanova-Páez, A. L. Montero-Alejo, M. A. Flores and W. Orellana, *Phys. B*, 2019, **568**, 81–87.
8 F. K. Alfadhili, A. B. Phillips, G. K. Liyanage, J. M. Gibbs, M. K. Jamarkattel and M. J. Heben, *MRS Adv.*, 2019, **4**, 913–919.
9 O. de Melo, M. Behar, J. F. Dias, R. Ribeiro-Andrade, M. da Silva, A. G. de Oliveira and J. C. González, *Mater. Sci. Semicond. Process.*, 2019, **97**, 17–20.
10 K. Luo, W. Wu, S. Xie, Y. Jiang, S. Liao and D. Qin, *Appl. Sci.*, 2019, **9**, 1885.
11 W.-C. Chen, C.-Y. Chen, Y.-R. Lin, J.-K. Chang, C.-H. Chen, Y.-P. Chiu, N.-I. Wu, K.-H. Chen and L.-C. Chen, *Interface engineering of CdS/CZTSSe heterojunctions for enhancing the* $Cu_2ZnSn(S,Se)_4$ *solar cell efficiency*, 2019, https://www.sciencedirect.com/science/article/pii/S2468606919300097.
12 K. Shen, X. Wang, Y. Zhang, H. Zhu, Z. Chen, C. Huang and Y. Mai, *Sol. Energy*, 2020, **201**, 55–62.
13 J. Miao, X. Liu, K. Jo, K. He, R. Saxena, B. Song, H. Zhang, J. He, M. Han, W. Hu and D. Jariwala, *Nano Lett.*, 2020, **20**, 2907–2915.
14 A. G. García and S. Zarate, *Microsc. Microanal.*, 2020, **26**, 2804–2805.
15 X. Zheng, E. Colegrove, J. N. Duenow, J. Moseley and W. K. Metzger, *J. Appl. Phys.*, 2020, **128**, 053102.
16 X. Yang, Y. Long, Y. Zheng, J. Wang, B. Zhou, S. Xie, B. Li, J. Zhang, X. Hao, S. Karazhanov, G. Zeng and L. Feng, *Mater. Sci. Semicond. Process.*, 2023, **156**, 107267.
17 C. Freysoldt, B. Grabowski, T. Hickel, J. Neugebauer, G. Kresse, A. Janotti and C. G. Van de Walle, *Rev. Mod. Phys.*, 2014, **86**, 253–305.
18 A. M. Ganose, D. O. Scanlon, A. Walsh and R. L. Z. Hoye, *Nat. Commun.*, 2022, **13**, 4715.
19 A. Mannodi-Kanakkithodi, The devil is in the defects, *Nat. Phys.*, 2023, **19**, 1243–1244.
20 S. R. Kavanagh, A. Walsh and D. O. Scanlon, *ACS Energy Lett.*, 2021, **6**, 1392–1398.
21 M. E. Turiansky, A. Alkauskas, M. Engel, G. Kresse, D. Wickramaratne, J.-X. Shen, C. E. Dreyer and C. G. Van de Walle, *Comput. Phys. Commun.*, 2021, **267**, 108056.
22 A. Wardak, W. Chromiński, A. Reszka, D. Kochanowska, M. Witkowska-Baran, M. Lewandowska and A. Mycielski, *J. Alloys Compd.*, 2021, **874**, 159941.
23 P. D. Hatton, M. J. Watts, Y. Zhou, R. Smith and P. Goddard, *J. Phys.: Condens. Matter*, 2022, **35**, 75702.
24 M. A. Scarpulla, B. E. McCandless, A. B. Phillips, Y. Yan, M. J. Heben, C. A. Wolden, G. Xiong, W. K. Metzger, D. Mao, D. Krasikov, I. Sankin, S. Grover, A. Munshi, W. Sampath, J. R. Sites, A. Bothwell, D. S. Albin, M. O. Reese, A. Romeo, M. Nardone, R. F. Klie, J. M. Walls, T. Fiducia, A. Abbas and S. M. Hayes, *Sol. Energy Mater. Sol. Cells*, 2023, **255**, 112289.
25 J.-H. Yang, W.-J. Yin, J.-S. Park, W. Metzger and S.-H. Wei, *J. Appl. Phys.*, 2016, **119**, 045104.

26 D. N. Krasikov, A. V. Scherbinin, A. A. Knizhnik, A. N. Vasiliev, B. V. Potapkin and T. J. Sommerer, *J. Appl. Phys.*, 2016, **119**, 085706.
27 D. Krasikov, A. Knizhnik, B. Potapkin, S. Selezneva and T. Sommerer, *Thin Solid Films*, 2013, **535**, 322–325.
28 T. Ablekim, S. K. Swain, W.-J. Yin, K. Zaunbrecher, J. Burst, T. M. Barnes, D. Kuciauskas, S.-H. Wei and K. G. Lynn, *Sci. Rep.*, 2017, **7**, 4563.
29 T. A. M. Fiducia, B. G. Mendis, K. Li, C. R. M. Grovenor, A. H. Munshi, K. Barth, W. S. Sampath, L. D. Wright, A. Abbas, J. W. Bowers and J. M. Walls, *Nat. Energy*, 2019, **4**, 504–511.
30 P. Gorai, D. Krasikov, S. Grover, G. Xiong, W. K. Metzger and V. Stevanović, *Sci. Adv.*, 2023, **9**, eade3761.
31 R. De Souza and G. Harrington, *Nat. Mater.*, 2023, **22**, 794–797.
32 D. V. Lang, *J. Appl. Phys.*, 2003, **45**, 3023–3032.
33 D. Wickramaratne, C. E. Dreyer, B. Monserrat, J.-X. Shen, J. L. Lyons, A. Alkauskas and C. G. Van de Walle, *Appl. Phys. Lett.*, 2018, **113**, 192106.
34 J. Y. Kim, L. Gelczuk, M. P. Polak, D. Hlushchenko, D. Morgan, R. Kudrawiec and I. Szlufarska, *npj 2D Mater. Appl.*, 2022, **6**, 75.
35 D. Broberg and K. Bystrom, *et al.*, *npj Comput. Mater.*, 2023, **9**, 72.
36 M. Y. Toriyama, J. Qu, G. J. Snyder and P. Gorai, *J. Mater. Chem. A*, 2021, **9**, 20685–20694.
37 C.-W. Lee, N. U. Din, K. Yazawa, W. Nemeth, R. W. Smaha, N. M. Haegel and P. Gorai, *J. Appl. Phys.*, 2024, **135**, 155101.
38 R. Grill and A. Zappettini, *Prog. Cryst. Growth Charact. Mater.*, 2004, **48–49**, 209–244.
39 J. Buckeridge, *Comput. Phys. Commun.*, 2019, **244**, 329–342.
40 A. Mannodi-Kanakkithodi, X. Xiang, L. Jacoby, R. Biegaj, S. T. Dunham, D. R. Gamelin and M. K. Y. Chan, *Patterns*, 2022, **3**, 100450.
41 A. Mannodi-Kanakkithodi, J.-S. Park, A. B. F. Martinson and M. K. Y. Chan, *J. Phys. Chem. C*, 2020, **124**, 16729–16738.
42 S. Kim, J.-S. Park, S. Hood and A. Walsh, *J. Mater. Chem. A*, 2019, **7**, 2686–2693.
43 J. L. Lyons and C. G. Van de Walle, *npj Comput. Mater.*, 2017, **3**, 1–10.
44 I. Mosquera-Lois, S. R. Kavanagh, A. Walsh and D. O. Scanlon, *npj Comput. Mater.*, 2023, **9**, 25.
45 I. Mosquera-Lois, S. R. Kavanagh, A. Walsh and D. O. Scanlon, *J. Open Source Software*, 2022, **7**, 4817.
46 M. P. Polak, R. Jacobs, A. Mannodi-Kanakkithodi, M. K. Y. Chan and D. Morgan, *J. Chem. Phys.*, 2022, **156**, 114110.
47 A. Mannodi-Kanakkithodi and M. K. Y. Chan, *J. Mater. Sci.*, 2022, **57**, 10736–10754.
48 T. Xie and J. C. Grossman, *Phys. Rev. Lett.*, 2018, **120**, 145301.
49 K. Choudhary and B. G. Sumpter, *AIP Adv.*, 2023, **13**, 095109.
50 Z. Chen, X. Li and J. Bruna, Supervised Community Detection with Line Graph Neural Networks, *arXiv*, 2017, preprint, arXiv:1705.08415v6, DOI: 10.48550/arXiv.1705.08415v6.
51 T. N. Kipf and M. Welling, Semi-Supervised Classification with Graph Convolutional Networks, *arXiv*, 2016, preprint, arXiv:1609.02907v4, DOI: 10.48550/arXiv.1609.02907v4.
52 M. D. Witman, A. Goyal, T. Ogitsu, A. H. McDaniel and S. Lany, *Nat. Comput. Sci.*, 2023, **3**, 675–686.
53 V. Fung, J. Zhang, E. Juarez and B. G. Sumpter, *npj Comput. Mater.*, 2021, **7**, 84.
54 G. Cheng, X.-G. Gong and W.-J. Yin, *Nat. Commun.*, 2022, **13**, 1492.
55 C. Chen and S. P. Ong, *Nat. Comput. Sci.*, 2022, **2**, 718–728.
56 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chem. Mater.*, 2019, **31**, 3564–3572.
57 J. Cheng, C. Zhang and L. Dong, *Commun. Mater.*, 2021, **2**, 92.
58 M. H. Rahman, P. Gollapalli, P. Manganaris, S. K. Yadav, G. Pilania, B. DeCost, K. Choudhary and A. Mannodi-Kanakkithodi, *APL Mach. Learn.*, 2024, **2**, 016122.
59 J. Lee and R. Asahi, *Comput. Mater. Sci.*, 2021, **190**, 110314.
60 M. H. Rahman, S. Rojsatien, D. Krasikov, M. K. Chan, M. Bertoni and A. Mannodi-Kanakkithodi, *Sol. Energy Mater. Sol. Cells*, 2025, **293**, 113857.
61 M. H. Rahman and A. Mannodi-Kanakkithodi, *J. Phys. Mater.*, 2025, **8**, 022001.
62 J. Pan, W. K. Metzger and S. Lany, *Phys. Rev. B*, 2018, **98**, 054108.
63 P. Borlido, J. Schmidt, A. W. Huran, F. Tran, M. A. L. Marques and S. Botti, *npj Comput. Mater.*, 2020, **6**, 98.
64 C. Vona, D. Nabok and C. Draxl, *Adv. Theory Simul.*, 2022, **5**, 2100496.
65 J. L. Lyons and C. G. Van de Walle, *npj Comput. Mater.*, 2017, **3**, 12.
66 A. Mannodi-Kanakkithodi, *Modell. Simul. Mater. Sci. Eng.*, 2022, **30**, 044001.
67 V. Bapst, T. Keck, A. Grabska-Barwińska, C. Donner, E. D. Cubuk, S. S. Schoenholz, A. Obika, A. W. R. Nelson, T. Back, D. Hassabis and P. Kohli, *Nat. Phys.*, 2020, **16**, 448–454.
68 C. Li, J. Poplawsky, Y. Yan and S. J. Pennycook, *Mater. Sci. Semicond. Process.*, 2017, **65**, 64–76.
69 C. Battaglia, A. Cuevas and S. De Wolf, *Energy Environ. Sci.*, 2016, **9**, 1552–1576.
70 D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue and T. Lookman, *Nat. Commun.*, 2016, **7**, 11241.
71 A. G. Kusne, T. Gao, A. Mehta, L. Ke, M. C. Nguyen, K.-M. Ho, V. Antropov, C.-Z. Wang, M. J. Kramer, C. Long and I. Takeuchi, *Sci. Rep.*, 2014, **4**, 6367.
72 F. Ren, L. Ward, T. Williams, K. J. Laws, C. Wolverton, J. Hattrick-Simpers and A. Mehta, *Sci. Adv.*, 2018, **4**, eaaq1566.
73 C. Kim, A. Chandrasekaran, A. Jha and R. Ramprasad, *MRS Commun.*, 2019, **9**, 860–866.
74 J. C. Verduzco, E. E. Marinero and A. Strachan, *Integrating Mater. Manuf. Innovation*, 2021, **10**, 299–310.
75 J. E. Gentle, *International Encyclopedia of Education*, Elsevier, 3rd edn, 2010, pp. 93–97.

76 J. Heyd, G. E. Scuseria and M. Ernzerhof, *J. Chem. Phys.*, 2003, **118**, 8207–8215.
77 M. H. Rahman and A. K. M. Kanakkithodi, *DefectDB: An Open Source Infrastructure for Defect Thermodynamics in II–VI Semiconductors*, 2025, https://nanohub.org/tools/defectdatabase.
78 A. Mannodi-Kanakkithodi, M. Y. Toriyama, F. G. Sen, M. J. Davis, R. F. Klie and M. K. Y. Chan, *npj Comput. Mater.*, 2020, **6**, 39.
79 G. Kresse and J. Furthmüller, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 11169–11186.
80 K. Choudhary and B. DeCost, *npj Comput. Mater.*, 2022, **8**, 221.
81 K. Choudhary, B. DeCost, L. Major, K. Butler, J. Thiyagalingam and F. Tavazza, *Digital Discovery*, 2023, **2**, 346–355.
82 D. E. Farache, J. C. Verduzco, Z. D. McClure, S. Desai and A. Strachan, *Comput. Mater. Sci.*, 2022, **209**, 111386.
83 I. Mosquera-Lois, S. R. Kavanagh, A. M. Ganose and A. Walsh, *npj Comput. Mater.*, 2024, **10**, 121.
84 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.
85 I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács and J. Riebesell, *J. Chem. Phys.*, 2024, **163**, 184110.
86 B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel and G. Ceder, *Nat. Mach. Intell.*, 2023, **5**, 1031–1041.
87 F. G. Sen, A. Mannodi-Kanakkithodi, T. Paulauskas, J. Guo, L. Wang, A. Rockett, M. J. Kim, R. F. Klie and M. K. Chan, *Sol. Energy Mater. Sol. Cells*, 2021, **232**, 111279.
88 J. Guo, A. Mannodi-Kanakkithodi, F. G. Sen, E. Schwenker, E. S. Barnard, A. Munshi, W. Sampath, M. K. Y. Chan and R. F. Klie, *Appl. Phys. Lett.*, 2019, **115**, 153901.
89 C. Freysoldt, J. Neugebauer and C. G. Van de Walle, *Phys. Rev. Lett.*, 2009, **102**, 016402.
90 L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, *Comput. Phys. Commun.*, 2020, **247**, 106949.
91 J. Laakso, L. Himanen, H. Homm, E. V. Morooka, M. O. Jäger, M. Todorović and P. Rinke, *J. Chem. Phys.*, 2023, **158**, 158.
92 M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos and E. Tuzhilina, *Nat. Rev. Methods Primers*, 2022, **2**, 100.
93 A. Zunger, S.-H. Wei, L. G. Ferreira and J. E. Bernard, *Phys. Rev. Lett.*, 1990, **65**, 353–356.
94 S. R. Kavanagh, A. G. Squires, A. Nicolson, I. Mosquera-Lois, A. M. Ganose, B. Zhu, K. Brlec, A. Walsh and D. O. Scanlon, *J. Open Source Software*, 2024, **9**, 6433.
95 P. Ščajev, M. Nardone, C. Reich, R. Farshchi, K. McReynolds, D. Krasikov and D. Kuciauskas, *Adv. Energy Mater.*, 2024, 2403902.
96 M. Tenorio, M. H. Rahman, A. Mannodi-Kanakkithodi and J. Chapman, *Chem. Phys. Rev.*, 2026, **7**, 011317.
97 F. A. Stevie and C. L. Donley, *J. Vac. Sci. Technol., A*, 2020, **38**, 063204.
98 J. Mahoney, C. A. Monroe, A. M. Swartley, M. G. Ucak-Astarlioglu and C. A. Zoto, *Spectrosc. Lett.*, 2020, **53**, 726–736.
99 A. Born, F. O. L. Johansson, T. Leitner, D. Kühn, A. Lindblad, N. Mårtensson and A. Föhlisch, *Sci. Rep.*, 2021, **11**, 16596.
100 G. Lanza, M. J. Jimenez, F. Alvarez, J. Pérez and A. Ávila, *ACS Omega*, 2022, **7**, 34521–34527.
101 H. Chen, D. T. L. Alexander and C. Hébert, *Nano Lett.*, 2024, **24**, 10177–10185.
102 H. Xie, X. Cheng and H. Huang, Investigation on the Interfaces in Organic Devices by Photoemission Spectroscopy, *Nanomaterials*, 2025, **15**, 680.
103 J. H. Lee, J. H. Lee, S. H. Jung, T. K. Hyun, M. Feng, J.-Y. Kim, J. Lee, H.-Y. Lee, J. S. Kim, C. Kang, K.-Y. Kwon and J. H. Jung, *Chem. Commun.*, 2015, **51**, 7463–7465.
104 L. Zhao, C. Sun, G. Tian and Q. Pang, *J. Colloid Interface Sci.*, 2017, **502**, 1–7.
105 Y. Li, G. Zha, D. Wei, F. Yang, J. Dong, S. Xi, L. Xu and W. Jie, *Sensors*, 2020, **20**, 2032.
106 T. Li, Y. Zhu, X. Ji, W. Zheng, Z. Lin, X. Lu and F. Huang, *J. Phys. Chem. Lett.*, 2020, **11**, 8901–8907.
107 D. Dragoni, T. D. Daff, G. Csányi and N. Marzari, *Phys. Rev. Mater.*, 2018, **2**, 013808.
108 E. Berger, M. Bagheri and H. Komsa, *Small*, 2025, **21**, e03956.
109 M. H. Rahman, E. H. Chowdhury and S. Hong, *Results Mater.*, 2021, **10**, 100191.
110 E. H. Chowdhury, M. H. Rahman and S. Hong, *Comput. Mater. Sci.*, 2021, **197**, 110580.
111 M. H. Rahman, M. Biswas and A. Mannodi-Kanakkithodi, *ACS Mater. Au*, 2024, **4**, 557–573.
112 M. H. Rahman, E. H. Chowdhury, D. A. Redwan, S. Mitra and S. Hong, *Phys. Chem. Chem. Phys.*, 2021, **23**, 5244–5253.
113 S. Mitra, M. H. Rahman, M. Motalab, T. Rakib and P. Bose, *RSC Adv.*, 2021, **11**, 30705–30718.
114 M. H. Rahman, Y. Sun and A. Mannodi-Kanakkithodi, *Mater. Adv.*, 2024, **5**, 8673–8683.

