

DeFecT-FF: Accelerated Modeling of Defects in Cd/Zn-Te/Se/S Compounds Combining High-Throughput DFT and Machine Learning Force Fields

Md Habibur Rahman^a and Arun Mannodi-Kanakkithodi^a

Abstract

We developed a framework for predicting the energies and ground state configurations of native point defects, extrinsic dopants and impurities, and defect complexes across zincblende-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 datasets of structures and energies from GGA-PBE and HSE06 geometry optimization, including bulk and alloyed supercells with and without defects, and defect-containing interface and grain boundary structures. Datasets 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 a nanoHUB tool, 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.

1 Introduction

Advancements in solar cell technologies are vital for meeting growing global energy demands and facilitating the transition to a decarbonized energy grid^{1,2}. Among available photovoltaic (PV) technologies, CdTe ranks as the second most widely used behind crystalline Si, accounting for approximately 7% of the global market^{1–3}. The commercial viability of CdTe in solar cells is primarily due to its direct bandgap of around 1.5 eV, which is well-suited for single-junction solar absorbers, along with its high absorption coefficient in the visible spectrum ($> 5 \times 10^5 \text{ cm}^{-1}$)⁴, low production costs, and favorable thin-film conductivity⁵. However, the highest recorded efficiency for CdTe solar cells is 22.3%, which remains below the theoretical limit of $\sim 30\%$, indicating considerable potential for improvement.

Shockley-Read-Hall (SRH) recombination is a major limiting factor that reduces the power conversion efficiency and is often associated with grain boundaries, point defects, and dislocations⁶. Native point defects and impurities can introduce energy levels or traps within the bandgap that act as nonradiative recombination

centers^{7,8} and reduce carrier lifetimes. E.g., studies have suggested that a Cd vacancy (V_{Cd}) in the CdTe lattice could significantly accelerate carrier recombination, leading to a reduction in carrier concentration and a decrease in power conversion efficiency by nearly 5%^{9,10}.

CdTe often suffers from low hole density, limiting its efficiency in PV applications^{1,2}. To mitigate this, Cu is typically introduced as an acceptor dopant, enhancing hole density^{1,2}. Cu doping is typically achieved through high-temperature annealing (673–723 K) in the presence of Cl, followed by a lower-temperature activation step (473–573 K)¹¹. During CdCl₂ treatment, Cl and Cu diffuse into CdTe at concentrations of 10^{17} – 10^{19} cm^{-3} , significantly altering electronic properties^{1,2}. While Cu_i and Cl_{Te} act as shallow donors, Cu_{Cd} behaves as a non-shallow acceptor, even forming defect complexes such as (Cu_i+Cu_{Cd}), (Cl_i+Te_{Cd}), and (Cl_i+Cu_{Cd})²⁺^{3,12–16}. However, Cu doping typically results in a suboptimal hole density (10^{14} cm^{-3}), far lower than the ideal 10^{16} cm^{-3} ⁵.

In contrast, group V doping (mainly As) achieves two to three times higher hole density without sacrificing carrier lifetime or photocurrent^{3,12–17}. Replacing the traditional CdTe absorber layer with Se-alloyed CdSe_xTe_{1-x} significantly improves CdTe PV efficiency by enhancing

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

long-wavelength absorption, reducing bandgap, increasing short-circuit current density, and improving carrier lifetimes due to better band alignment and reduced interfacial recombination^{5,17,18}. Se passivates grain boundaries and dislocation cores, further extending carrier lifetimes. ZnTe is widely employed as a hole transport layer due to its favorable band alignment with CdTe, facilitating efficient hole collection and enhancing device performance¹⁹. Given these factors, exploring the defect chemical space of Cd/Zn-S/Se/Te bulk and alloyed compositions and interfaces in terms of single native and extrinsic point defects as well as defect complexes, is essential to understanding defect-driven solar cell efficiency and ultimately guiding the development of more efficient and stable CdTe- and CdSeTe-based thin-film PV devices²⁰.

Defect levels are typically measured experimentally using cathodoluminescence, photoluminescence, optical spectroscopy, or deep-level transient spectroscopy (DLTS)²¹. These methods can be challenging due to difficulties in sample preparation and assigning measured levels to specific defects²². For instance, distinguishing whether an observed peak originates from a particular vacancy, self-interstitial, or an unintended substitutional or interstitial impurity, or a combination of them, is not straightforward²³. To address these challenges, density functional theory (DFT) computations have become a widely used approach for determining the formation energy (E_f) of point defects as a function of Fermi level (E_F), the residual charge in the system (q), and chemical potential conditions (μ)^{6,24-28}.

DFT enables the reliable identification of the lowest energy donor- and acceptor-type defects in a solid, the range of possible shallow and deep defect levels, the equilibrium conductivity as determined by the most stable defects (p-type, intrinsic, or n-type), defect concentrations as a function of temperature, electron/hole capture rates, and many other related properties^{10,29-36}. When an appropriate level of theory is applied, DFT-computed charge transition levels often show excellent agreement with experimentally measured levels^{23,29}. However, DFT simulations are time-consuming and scale poorly with the number of atoms, causing major issues in using large supercells to represent point defects. It is especially a huge challenge to use DFT for exploring a broad configurational space of potential vacancies, interstitials, antisite defects and defect complexes across dozens of compounds or chemistries of interest³⁷. Any given defect can adopt numerous configurations and charge states, further adding to the cost of exhaustive exploration³⁸.

The prediction of defect properties can be greatly accelerated by integrating DFT simulations with state-of-the-art machine learning (ML) approaches such as crystal graph neural networks (GNNs)³⁹⁻⁴¹. GNNs have gained immense popularity in recent years and are now widely used to effectively represent and predict energies and properties of molecules, polymers, and solid-state crystalline materials^{42,43}. They operate on graph-structured data by transforming crystal structures into crystal graphs where atomic positions are represented as nodes and chemical bonds as edges⁴⁴. They are capable of learning intricate internal representations within crystalline environments, which are valuable for predicting various properties of interest such as formation or decomposition energy, electronic bandgap, and defect formation energy, while significantly reducing computational costs.

In past work, we demonstrated the use of GNNs to accelerate the prediction and screening of native defects and functional impurities in the chemical space of group IV, III-V, and II-VI zincblende (ZB) semiconductors⁴⁵. This dataset encompasses a wide range of defect types, including vacancies, self-interstitials, anti-site substitutions, and extrinsic interstitial and substitutional defects. While the models were capable of predicting charge-dependent defect formation energies for cation-rich and anion-rich chemical potential conditions directly from an assumed defect structure, they had several limitations: (1) They were trained on a broad chemical space (34 compounds, and practically any element from across the periodic table as a possible defect), which led to large errors on specific compositions and combinations of defects. (2) Models worked well for single native or extrinsic defects in binary compounds, but did not show the same accuracy for ternary or quaternary compositions often encountered in CdTe solar cells (e.g., Se rich CdSe_xTe_{1-x} local environment and Cd_xZn_{1-x}Te local environment between CdTe and ZnTe interface), or defect complexes in a variety of compositions^{2,11,13,31,46-48}. (3) They were primarily trained on data from cubic 2 × 2 × 2 supercells with singular point defects, thus affecting their applicability to larger supercells necessary for defect complexes. (4) They were trained on defect configurations from the cheaper PBE (Perdew–Burke–Ernzerhof) functional within the generalized gradient approximation (GGA), rather than a hybrid functional or other high-fidelity method⁴⁹⁻⁵¹. (5) For obtaining ground state defect configurations, the models must be combined with gradient-free stochastic optimization techniques, preventing them from performing more efficient gradient-based geometry optimization.

Additionally, interfaces—such as those between CdTe (or CdSeTe) and back contact ZnTe—are very important in devices where compositional grading and lattice mismatch can strongly influence dopant incorporation, defect segregation, and carrier recombination^{52–60}. Likewise, grain boundaries are inherent to polycrystalline CdTe and often serve as preferential sites for defect clustering, impurity segregation, or charge trapping^{61–63}. Understanding point defect energetics in these extended defect structures is therefore critical for improving device stability and performance^{61,62,64,65}. Our previous defect GNN models did not account for the presence of extended structural features, further underscoring the necessity of extending them to more complex compositions, larger supercells, defect complexes, and higher levels of theoretical accuracy. By incorporating interface and grain boundary configurations into the training data, the model can generalize better to real-world polycrystalline and heterostructured devices, offering predictive insights for defect engineering beyond idealized bulk conditions^{66–70}.

The accuracy of ML models is heavily dependent on the quantity and quality of the training data. In the present context, the DFT data for defect structures must encompass all regions of the desired chemical space and all types of defects. By using active learning (AL), new DFT data can be systematically and rationally generated at every step of GNN model retraining by picking new computations that reduce the prediction uncertainty^{71–77}. In this work, we used AL to perform new defect calculations to gradually improve the GNN models and ultimately accurately identify low-energy defect configurations in the chemical space of Cd/Zn-Te/Se/S compositions. This specific chemical space is chosen because: (a) Se grading is used in CdTe solar cells, and it is important to predict defect properties as a function of Se concentration, (b) the absorber layer forms interfaces with back contacts such as ZnTe, and Cd-Zn composition grading might thus be important, (c) by tuning the absorber composition itself, defect properties could be manipulated, and (d) examining all low energy defects and dopants across Cd/Zn-Te/Se/S compositions will provide a better library for comparison with experimental measurements, which may arise from different sorts of local coordination environments and chemistries, including defect complexes.

While compositions such as ZnS and ZnSe may not directly be of interest to CdTe-based solar cells, they are still chemically informative to the GNN models, and the predicted properties may be useful for understanding

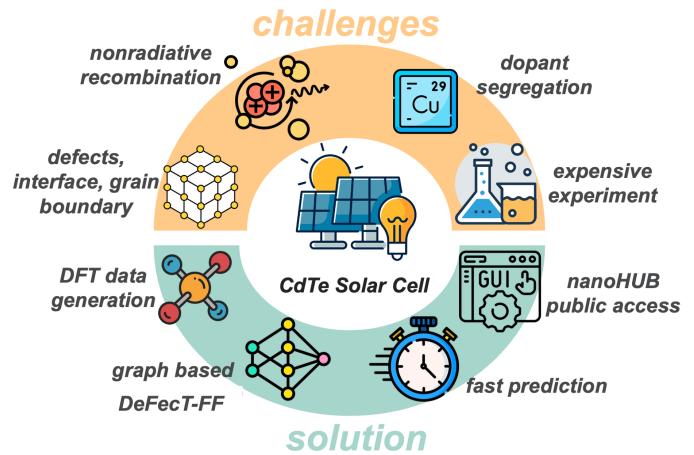


Figure 1 Key challenges limiting CdTe solar-cell performance—point defects, interfaces and grain boundaries, nonradiative carrier recombination, dopant segregation, and the high cost of experiments—and our four-step solution: (i) high-throughput DFT data generation, (ii) machine learning force field (MLFF) training, (iii) rapid large-scale prediction, and (iv) open, public access for defect modeling in $\text{CdSe}_x\text{Te}_{1-x}$ solar cells.

local coordination environments that may exist in the solar cell where various components include Cd, Zn, Te, Se, and S. CdS is important because of its use as a buffer layer, though it will likely not be directly relevant as a solar absorber. Interfaces between many of these chalcogenide compounds, which primarily adopt the ZB phase, are certainly relevant to the performance of CdTe and $\text{CdSe}_x\text{Te}_{1-x}$ solar cells. Our goal is to develop a model capable of making predictions on suitably large supercells and applicable to defect complexes and grain boundaries in mixed compositions and interfaces. Furthermore, it is vital to extend our predictive capabilities to the non-local hybrid HSE06 (Heyd-Scuseria-Ernzerhof) functional which improves upon the many limitations of semi-local DFT⁷⁸.

Our methodology begins with an initial dataset containing bulk and defect configurations in Cd/Zn-Te/Se/S binary and multi-nary compounds, which was used to train early GNN models at GGA-PBE accuracy⁷⁹. These models served as the foundation for predicting defect properties across the entire hypothetical defect chemical space that includes thousands of vacancies, interstitials, anti-site defects, substitutional defects, and defect complexes, considering all possible native defects, group V dopants, and unintentional impurities such as Cl and O, which are relevant to CdTe and $\text{CdSe}_x\text{Te}_{1-x}$ solar cells^{80,81}. To efficiently guide the discovery of new low energy defects, we used Bayesian optimization-based AL, where the next set of DFT simulations was selected based on an acquisition

function that aimed to minimize the prediction uncertainty.

Figure 1 captures some of the major challenges in CdTe solar cell performance and our proposed approach for tackling some of them in terms of better defect predictions. Finally, leveraging the augmented and curated HSE06 dataset built upon thousands of bulk and defect configurations optimized with GGA-PBE, we trained a GNN-based machine learning force field (MLFF) model (packaged as “DeFecT-FF”) using DFT-derived *energies, atomic forces, and stresses*. This MLFF is based on the M3GNet framework⁴⁴ and once rigorously trained, is capable of optimizing both single defects and defect complexes within the entire Cd/Zn-Te/Se/S chemical space, at both PBE and HSE06 accuracy. The next few sections present details of our methodology and a discussion of the results, ending with some important case studies involving defect formation energy plots from HSE06 with spin-orbit coupling (SOC) that demonstrate the effectiveness and generalizability of the DFT-MLFF approach, and the significant reduction in DFT computational time for defect prediction and screening.

2 Description of the DFT datasets

For the GGA-PBE dataset, we collected a majority of the structures and energies from multiple past published works from our group^{29,45,82,83}, and combined them with some systematic new computations on previously unexplored alloy compositions and defect structures. Upwards of 20,000 bulk and defect configurations were part of this dataset, along with their total energies, atomic forces, and stresses collected from the DFT geometry optimization trajectories. Starting from selected bulk and defect configurations from the PBE dataset, HSE06 geometry optimization was then performed using a default mixing parameter of $\alpha = 0.25$. For defect calculations, lattice parameters were changed from the PBE-optimized values to the correct HSE06-optimized values to enable relaxation with fixed volume.

Combining the bulk structures (binary, ternary, and quaternary compounds; specific alloy series are listed in **Figure 2(a)**) and defect-containing structures into one dataset for training the GNN models helps simultaneously take into account the effects of alloying (e.g., Cd-Zn mixing or Se-Te mixing), cation and anion ordering, supercell size, and the type and distribution of point defects. Every structure is characterized by its “crystal formation energy” (CFE), which is simply the per-atom energy required to split any bulk or defect crystal into its constituent atoms.

GNN models can be trained to either directly predict the CFE or indirectly predict it following a force field-based geometry optimization. From the CFE, either the bulk stability (in terms of a formation energy, decomposition energy, or energy above hull) or the defect formation energy can subsequently be estimated by using known energies of relevant reference phases.

Our chemical space includes all possible Cd/Zn-Te/Se/S binary compounds as well as selected series of ternary and quaternary compounds, specifically $\text{CdSe}_x\text{Te}_{1-x}$, $\text{CdS}_x\text{Se}_{1-x}$, $\text{Cd}_x\text{Zn}_{1-x}\text{S}$, $\text{Cd}_x\text{Zn}_{1-x}\text{Se}$, $\text{Cd}_x\text{Zn}_{1-x}\text{Te}$, $\text{ZnSe}_x\text{Te}_{1-x}$, $\text{ZnS}_x\text{Se}_{1-x}$, $\text{Cd}_{0.5}\text{Zn}_{0.5}\text{S}_x\text{Se}_{1-x}$, and $\text{Cd}_{0.5}\text{Zn}_{0.5}\text{Se}_x\text{Te}_{1-x}$. The mixing fraction x systematically varies in multiples of 0.125 (or $n/8$ mixing, where n is a positive integer between 1 and 8), leading to a total of 81 unique compositions⁸². All such structures are simulated in both $2 \times 2 \times 2$ (64 atoms) and $3 \times 3 \times 3$ (216 atoms) cubic ZB supercells using the special quasirandom structures (SQS)⁸⁴ approach, starting from known binary compound ground state configurations. From across the set of all PBE calculations, a total of 10,080 bulk configurations were collected, including intermediate structures from the geometry optimization trajectories. The CFE is calculated as follows, taking $\text{CdSe}_x\text{Te}_{1-x}$ as an example:

$$\text{CFE} = \frac{E(\text{Cd}_a\text{Se}_b\text{Te}_c) - aE(\text{Cd}) - bE(\text{Se}) - cE(\text{Te})}{N_{\text{atoms}}} \quad (1)$$

Here, $E(\text{Cd}_a\text{Se}_b\text{Te}_c)$ is the total DFT energy of the supercell containing a atoms of Cd, b atoms of Se, and c atoms of Te that is necessary to simulate the $\text{CdSe}_x\text{Te}_{1-x}$ composition. $E(\text{Cd})$, $E(\text{Se})$, and $E(\text{Te})$ are respectively the per-atom energies of Cd, Se, and Te in their known elemental standard states, and $N_{\text{atoms}} = a+b+c$ is the total number of atoms in the supercell.

The dataset compiled from past publications contains a whole host of native defects (vacancies, self-interstitials, anti-site substitutions), extrinsic interstitials and substitutional defects accounting for both unintentional impurities and intentional doping, and a handful of defect complexes, in different Cd/Zn-Te/Se/S compositions. The number of $2 \times 2 \times 2$ supercell defect structures for different charge states (total charge q imposed on the defect simulation during geometry optimization) accounted for the following: 7302 ($q=+2$), 6201 ($q=+1$), 8203 ($q=0$), 6361 ($q=-1$), and 7689 ($q=-2$). We performed a series of new defect (native and extrinsic) calculations using $3 \times 3 \times 3$ supercell structures for some selected compounds such as CdTe and $\text{CdSe}_{0.25}\text{Te}_{0.75}$, as well as in CdTe-ZnTe

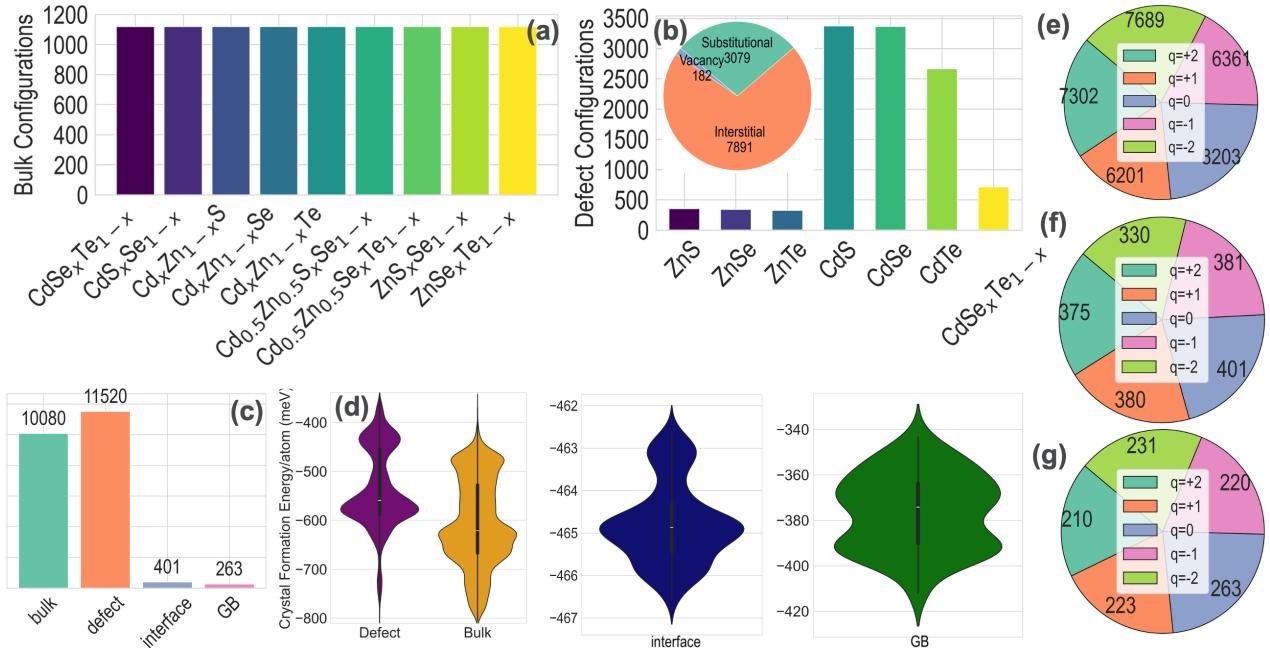


Figure 2 Statistics of the GGA-PBE dataset: (a) Number of bulk configurations corresponding to $\text{CdSe}_x\text{Te}_{1-x}$, $\text{CdS}_x\text{Se}_{1-x}$, $\text{Cd}_x\text{Zn}_{1-x}\text{S}$, $\text{Cd}_x\text{Zn}_{1-x}\text{Se}$, $\text{Cd}_x\text{Zn}_{1-x}\text{Te}$, $\text{Cd}_{0.5}\text{Zn}_{0.5}\text{S}_x\text{Se}_{1-x}$, $\text{Cd}_{0.5}\text{Zn}_{0.5}\text{Se}_x\text{Te}_{1-x}$, $\text{ZnS}_x\text{Se}_{1-x}$, and $\text{ZnSe}_x\text{Te}_{1-x}$ compositions. (b) Number of neutral defect configurations in CdS, CdSe, CdTe, ZnS, ZnSe, ZnTe, and different $\text{CdSe}_x\text{Te}_{1-x}$ compositions, with the inset showing the distribution of vacancy, substitutional, and interstitial defects across the dataset. (c) Bar chart showing the total number of bulk, defect, interface, and grain boundaries (GB) configurations in the dataset. (d) Violin plots showing the distribution of crystal formation energy across all the bulk, defect, interface and grain boundary (GB) structures. Inside each violin, a mini box plot shows the median (central line), quartile, and range (whiskers). (e, f, g) Distribution of defect configurations for different charge states ($q = +2, +1, 0, -1$, and -2) in bulk defect, interface, and GB configurations, respectively.

Dataset	Supercell Size	Data Points
Bulk dataset from Cd/Zn-Te/Se/S binary, ternary and quaternary alloys	$2 \times 2 \times 2$	10080
Bulk dataset from $\text{CdSe}_x\text{Te}_{1-x}$ alloys	$3 \times 3 \times 3$	26
Defect dataset from 6 Cd/Zn-Te/Se/S binary compounds	$2 \times 2 \times 2$	7302 ($q=+2$), 6201 ($q=+1$), 8203 ($q=0$), 6361 ($q=-1$), 7689 ($q=-2$)
Defect dataset from $\text{CdSe}_x\text{Te}_{1-x}$ and $\text{Cd}_x\text{Zn}_{1-x}\text{Te}$ alloys	$3 \times 3 \times 3$	10 ($q=+2$), 12 ($q=+1$), 26 ($q=0$), 18 ($q=-1$), 21 ($q=-2$)
Defect dataset from CdTe/ZnTe interface	$3 \times 3 \times 6$	375 ($q=+2$), 380 ($q=+1$), 401 ($q=0$), 381 ($q=-1$), 330 ($q=-2$)
Defect dataset from CdTe grain boundaries	–	210 ($q=+2$), 223 ($q=+1$), 263 ($q=0$), 220 ($q=-1$), 231 ($q=-2$)

Table 1 Number of data points (or structures) in the GGA-PBE dataset corresponding to different types of bulk or defect configurations, supercell sizes, and charge states.

heterostructures and CdTe grain boundary structures collected from other past work⁸⁵. Defect species included both native defects and extrinsic impurities and dopants of interest such as Cl, O, Cu, As, and P.

The CdTe-ZnTe interface was simulated by merging optimized $3 \times 3 \times 3$ supercells of CdTe and ZnTe with slightly strained lattice parameters, followed by volume-free geometry optimization. For all new defect calculations in $3 \times 3 \times 3$ supercell binary and alloyed structures, heterostructures, and grain boundaries, the Doped package⁸⁶ was used to generate a series of starting configurations based on bond distortions and atomic rattling, to adequately account for symmetry-breaking around defect centers. Ultimately, hundreds of new defect structures

were generated using larger bulk supercells and the interface and grain boundary structures, and merged with the $2 \times 2 \times 2$ supercell defect structure data compiled from past work.

In the compiled dataset, the label `bulk` denotes pristine supercells without any point defects, representing ideal defect-free crystal structures. The label `defect` refers to supercells containing a point defect, such as a vacancy, self- or extrinsic interstitial, anti-site or extrinsic substitution, or a defect complex combining two or three of these individual defects. The label `interface` refers to large heterostructures with a variety of point defects at the CdTe-ZnTe interface, capturing the influence of interfacial environments on defect behavior. The label `GB` designates

grain boundary structures containing native or extrinsic defects, which helps study the interplay between point defects and extended structural features. The statistics of the compiled PBE dataset corresponding to charge states $q = +2, +1, 0, -1$, and -2 are presented in **Figure 2(a-g)**, **Figure S1** and **Table 1**. **Figure 2(d)** shows violin plots capturing the distribution of CFEs ($q=0$) across the dataset for bulk and defect configurations (including interface and GB structures), with values ranging from ~ -800 meV/atom to ~ -320 meV/atom. Before GNN models were trained, energetically and configurationally similar structures were removed, leading to a reduced and curated dataset of approximately 6,923 bulk configurations and 5,910 ($q=+2$), 4,929 ($q=+1$), 6,242 ($q=0$), 4,795 ($q=-1$), and 5,666 ($q=-2$) defect configurations. In this work, we used two established GNN approaches to train predictive models on the DFT dataset: Atomistic Line Graph Neural Network (ALIGNN)⁸⁷, and the Materials 3-body Graph Network (M3GNet)-based machine learning force field (MLFF)⁴⁴.

3 Graph neural network models for direct prediction of crystal formation energy

ALIGNN was developed by Choudhary *et al.*⁸⁷ and considers both two-body (bond lengths) and three-body interactions (bond angles). ALIGNN leverages both graph convolution layers and line graph convolution layers to capture short-range and long-range correlations in the crystal. For training the ALIGNN models, the learning rate was set to 0.001, an AdamW optimizer was used to update the weights and biases of the model, 4 graph convolution layers and 4 line graph layers were implemented, the cutoff radius was set to 6 Å with 12 nearest neighbors to create the crystal graph, and models were trained up to 90 epochs with a batch size of 8. We experimented with different training-validation-test splits of the dataset and found that the 60:20:20 ratio works the best.

ALIGNN models were trained to predict the CFE from any given bulk or defect crystal structure, using only the neutral charge state structures at this stage. Parity plots capturing the performance of the optimized models are presented in **Figure 3**, in terms of ALIGNN-predicted CFE vs DFT-computed CFE for only the test set data points. Models pictured in **Figure 3(a-c)** are respectively trained only on bulk structures, only on defect structures, and on both bulk and defect structures; this distinction is made to understand how sensitive the models are to different types of configuration. As shown in **Figure 3(a)**, the ALIGNN

model for bulk structures alone shows a test prediction root mean squared error (RMSE) of 1.43 meV/atom, and this error remains practically unchanged for the combined model in **Figure 3(c)**. For the defect-only ALIGNN model in **Figure 3(b)**, test RMSE ranges from 3.09 meV/atom for interstitial defects to 4.87 meV/atom for substitutional defects to 8.36 meV/atom for vacancy defects. Each of these defect prediction errors comes down for the combined data model, proving the value of increasing the size and chemical and structural diversity of the training dataset.

The training dataset contains a larger number of interstitial defects, followed by substitutional and vacancy defects: this is mostly a consequence of there being a lot more options for interstitial and substitutional defects in terms of extrinsic species from across the periodic table, and also the longer time it takes for DFT optimization of these defect structures, leading to more intermediate geometries. This results in the comparatively higher RMSE for CFE prediction of vacancy defects (6.47 meV/atom) than substitutional (3.84 meV/atom) and interstitial (2.04 meV/atom) defects. The ALIGNN predictions for all types of structures are highly accurate with vanishingly small errors considering the total range of CFE values. GNN models for defect predictions reported in the recent literature^{43,45,88} primarily focused on sampling defect configurations to train surrogate models, while we have adopted the approach of combining bulk and defect structures which enhances the overall generalizability and accuracy of the models.

Although state-of-the-art GNNs are very robust and powerful for modeling complex relationships within atomic structures^{39,40,87,89-93}, their transferability beyond the trained dataset remains questionable. To evaluate this, we performed the following series of tests:

1. An ALIGNN model was trained purely on bulk structures and then used to predict the CFE of defect structures.
2. An ALIGNN model was trained only on interstitial defect structures and used to predict the CFE for vacancy and substitutional defects.
3. An ALIGNN model trained exclusively on defects in $2 \times 2 \times 2$ supercell structures and then used to predict the CFE of $3 \times 3 \times 3$ supercell defect structures.

ALIGNN shows poor transferability when trained only on bulk data and used to predict for defects; as shown

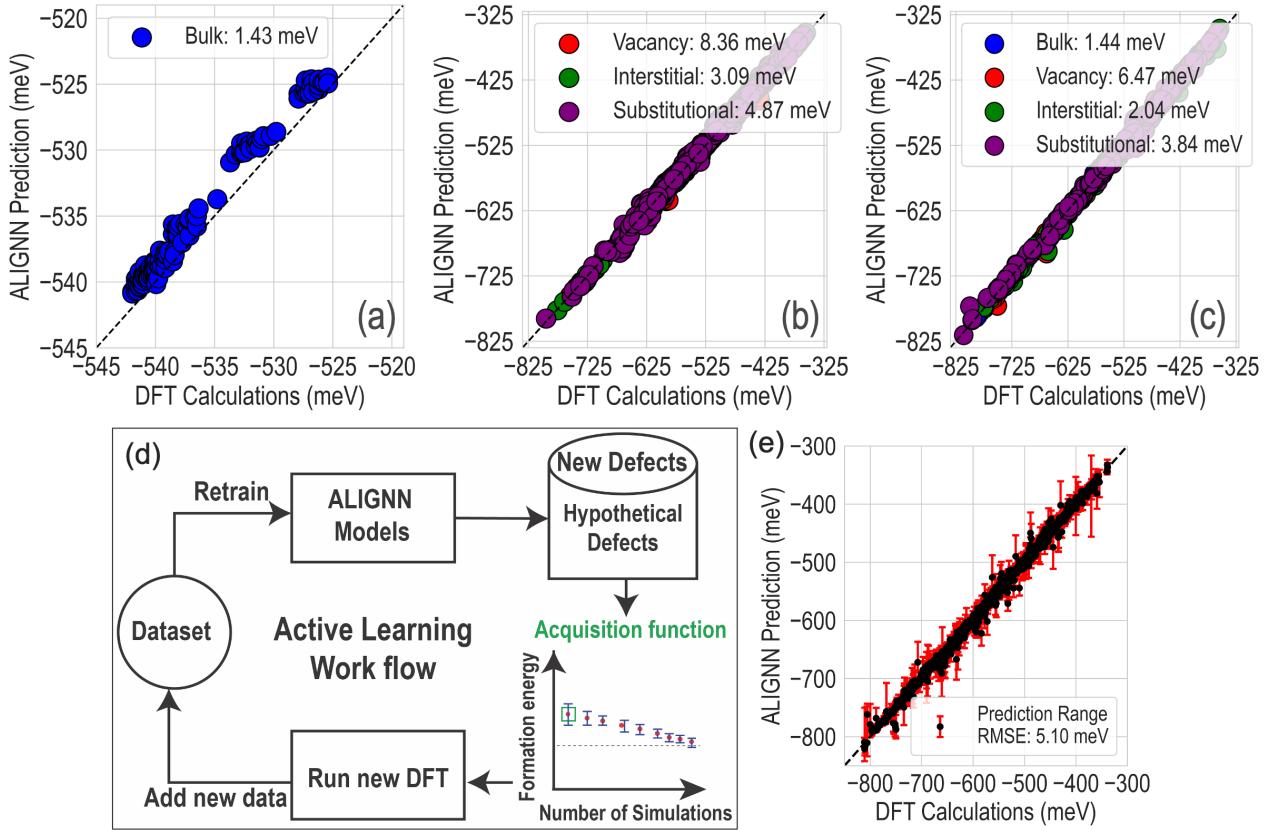


Figure 3 Parity plots for ALIGNN models trained on the GGA dataset using (a) only bulk structures, (b) only defect structures, and (c) both bulk and defect structures. (d) Active learning (AL) workflow implemented in this work: standard deviation of ALIGNN-predicted formation energy of novel defects is used to determine acquisition functions and identify the next set of DFT simulations to run. The ALIGNN model is then retrained with the new data and predictions are made for the remaining set of unexplored defects. (e) An ALIGNN vs DFT parity plot showing standard deviation in test set prediction across 100 separate models; these error bars are used to calculate the acquisition functions in the AL workflow.

in **Figure S2(a)**, the prediction RMSEs range from ~ 36 meV/atom for vacancies to > 50 meV/atom for interstitial and substitutional defects. Since this model has not been exposed to specific configurations such as atomic relaxation around defect sites, it is unable to predict as accurately for defect structures as it does for bulk. **Figure S2(a)** also shows that ALIGNN uniformly under-predicts the CFE of all defect configurations, which could be attributed to the lower average CFE values in the bulk dataset compared to defects, as illustrated by the violin plot in **Figure 2(e)**. The model trained only on interstitial defects does a reasonable job for vacancy defects, but the RMSE values are larger than from the models in **Figure 3** and there are some very clear outliers, as pictured in **Figure S2(b)**. Lastly, when the model is trained on only $2 \times 2 \times 2$ supercell structures and used to predict for $3 \times 3 \times 3$ supercells, the predictions show good accuracy but a slight tendency to over-estimate the CFE, hinting at the fact that ALIGNN may be capable of extrapolating across supercell sizes. These results suggest that the GNN models for CFE

prediction could generalize across types of chemistries, structures, and system sizes for particular cases, but in general may need to be retrained and fine-tuned for specific datasets.

4 Expanded defect chemical space

The GNN models are ultimately intended for prediction and screening across thousands of possible novel defect configurations. Considering a series of Cd/Zn-Te/Se/S compositions, SQS structures, interfaces between them, and all possible native defects, group V dopants, impurities such as Cl and O, etc., there are potentially several hundred thousand single defects and double or triple defect complexes possible. Here, we defined an expanded defect chemical space using a subset of this entire available space, to keep the numbers tractable and enable predictions at a reasonable rate. We considered 9 possible compounds, namely CdTe, CdSe, ZnTe, three Cd-Se-Te compositions, and three Cd-Zn-Te compositions, and populated possi-

ble native defects (vacancies, self-interstitials, and anti-site substitutions), selected extrinsic defects (Cu, P, As, Sb, Cl, and O at Cd/Zn/Te/Se or interstitial sites), and selected defect complexes which are combinations of native defects and extrinsic defects. This chemical space is pictured in **Figure S3** across the 9 compounds, showing a total of ≈ 25000 unique defect structures, accounting for both singular defects and defect complexes.

Symmetry-breaking is important for defect structures, and there could be many possible sites when considering defects in alloys or defect complexes. For all the defects described above, the `Doped`⁸⁶ package was used to identify possible defect sites in $3 \times 3 \times 3$ supercells of binary and ternary compounds. For instance, As_{Te} (As at Te site) in $\text{Cd}_{0.75}\text{Zn}_{0.25}\text{Te}$ has 108 non-equivalent Te sites, and all of them are taken into account in our expanded defect space. Theoretically, a non-symmetric supercell, such as any $\text{CdSe}_x\text{Te}_{1-x}$ compound, could host thousands of possible interstitial defects. Typically, interstitial defects are less stable than vacancies, substitutional defects, or anti-sites, as demonstrated in prior work⁴⁵. Therefore, we relaxed the criteria when generating interstitial defect configurations using `Doped`.

5 Active learning (AL) workflow with GNN-based geometry optimization

We employed an AL framework to efficiently explore the expanded defect space by iteratively training ensembles of 100 ALIGNN models each at different steps, before obtaining the best models pictured in **Figure 3(a-c)** and discussed earlier. The general AL workflow and results are pictured in **Figure 3(d-e)** and **Figure S4**. Mean and standard deviation of the CFE predictions were estimated using the ensemble of ALIGNN models. Based on a maximum uncertainty acquisition function, 200 of the most informative defect structures were selected for full DFT optimization and added to the training set in each AL cycle. **Figure 3(e)** shows an example model at one of the steps: plotted are ALIGNN-predicted CFE vs DFT-computed CFE (from subsequent calculations) for selected points from the expanded defect space, with error bars capturing the standard deviation in prediction across 100 models.

This iterative process rapidly improved model accuracy and predictive performance, even in relatively unexplored regions of the broader chemical space. Additional details of the AL workflow are presented in the SI. An important question we faced while making predictions for new defect structures is the following: how do we go beyond the CFE prediction for the initial defect configuration (including distorted structures generated from `Doped`) and determine

the ground state configuration corresponding to the given defect? To achieve this, the ALIGNN prediction must be coupled with an optimization algorithm that relaxes atoms in the neighborhood of a defect and yields the lowest-energy structure. A version of this was presented in our past work⁴⁵, where hundreds of configurations with random atomic distortions were generated for any defect structure and GNN-predicted energies were used to determine the configuration closest to ground state.

Here, we combined the ALIGNN predictions with two commonly used optimization techniques to achieve energy minimization of new defect configurations:

1. Simulated Annealing (SA), which is a gradient-free stochastic method employing random atomic perturbations guided by an annealing schedule.
2. Bayesian Optimization (BO), which uses Gaussian processes to iteratively identify low energy configurations through atomic displacement exploration.

Coupling ALIGNN prediction of CFE values with SA or BO enabled gradient-free energy minimization via atomic displacements applied within a cut-off radius around the defect center. These algorithms systematically searched for the lowest energy configuration using ALIGNN predictions at each step, allowing fast and guided optimization. As a case study, we applied both algorithms to optimize a Cd vacancy (V_{Cd}) defect in a $3 \times 3 \times 3$ CdTe supercell, and the results are pictured in **Figure S5**. Either method reaches a general energy convergence within 300 to 400 optimization steps which take a total of a few minutes to complete, but SA performs much better than BO: not only does it find a lower energy structure, but it actually discovers the configuration featuring a Te–Te dimer, which was reported by Kavanagh et al.⁹ and which is easily missed by standard DFT optimization. To further evaluate the optimization capability of the ALIGNN model, we combined it with SA to optimize a variety of defects across the $\text{CdSe}_x\text{Te}_{1-x}$ and $\text{Cd}_x\text{Zn}_{1-x}\text{Te}$ chemical space as shown in **Figure S6**. For example, Cl_{Te} in $\text{Cd}_{0.50}\text{Zn}_{0.50}\text{Te}$ was successfully optimized using ALIGNN+SA, as shown in **Figure S6 (a)**.

Figure S7 illustrates the DFEs of As_{Te} and Cl_{Te} in $\text{Cd}_{0.50}\text{Zn}_{0.50}\text{Te}$ under Cd-rich conditions as a function of E_F , computed using the HSE06+SOC functional on top of PBE-optimized lowest-energy sites. The right panel compares two workflows and highlights the computational advantage of incorporating ALIGNN+SA-based optimization. In the conventional DFT-only approach, all 108 symmetry-

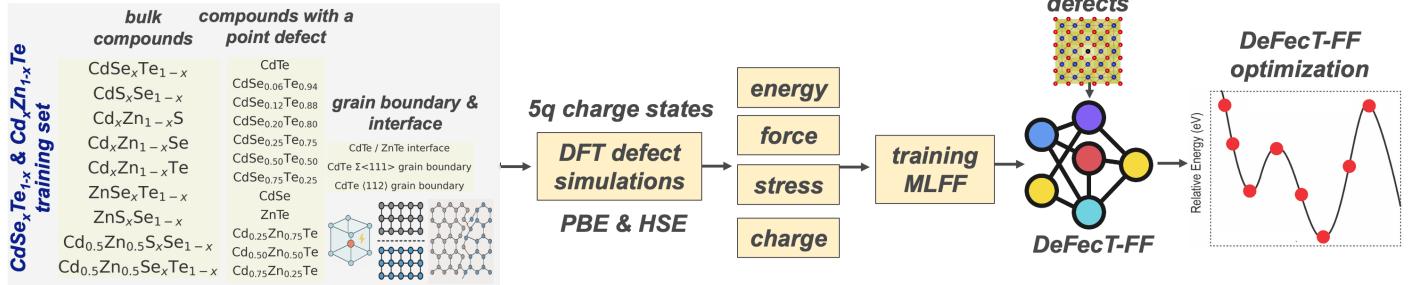


Figure 4 The DeFect-T-FF workflow for simulating and understanding defects in Cd/Zn-Te/Se/S compounds. The training set includes bulk binary, ternary, and quaternary compounds, point-defect structures, grain boundary structures with defects, and interface structures with defects, simulated in five charge states ($q = +2, +1, 0, -1, \text{ and } -2$) using both GGA-PBE and HSE06 functionals. Charge-specific MLFF models are trained to eventually enable geometry optimization of any new defect configurations.

inequivalent As_{Te} configurations must be relaxed individually to identify the lowest-energy structure, requiring approximately 512 hours. In contrast, the ALIGNN+DFT workflow first uses the trained ALIGNN+SA model to rapidly evaluate and optimize the CFE for all configurations, identifying the most stable site in minutes. DFT calculations (PBE followed by HSE06+SOC) are then performed on the ALIGNN-predicted lowest-energy structure, reducing the total computational cost to just 8 hours. The next section presents an attempt to improve upon this by moving towards gradient-based optimization using force fields.

6 Training crystal graph-based machine learning force field (MLFF)

M3GNet-based MLFF models were trained using the energies, atomic forces, and stresses extracted from GGA-PBE calculation trajectories. Figure 4 shows the workflow for training these models, eventually yielding the ability to optimize any new defect configuration and predict its ground state energy. Radial and three-body cutoffs were set to 6 Å and 6 Å, respectively. The loss was a weighted sum of RMSE for energies, forces, and stresses (weights = 1, 1, and 0.01, respectively). Training was performed on an NVIDIA A100 (80 GB) with batch size 64 and learning rate 5×10^{-4} until convergence. Geometry optimization with the MLFF used the FIRE algorithm in ASE^{94,95} with convergence criteria of mean atomic force $< 10^{-5}$ eV/Å or a maximum of 100 ionic steps. Models were trained separately for structures in five different charge states. To improve model accuracy on difficult configurations where predictions were poor, we used a two-stage training process:

- **Warm-up:** In the first stage, we trained the model for a small number of epochs using uniform sampling. This helps the model develop a basic understanding of the data.

- **Error-aware reweighting:** Next, we used the warm-up model to predict energies and forces for all training samples. Based on the prediction errors, we assigned a score to each sample. Samples with larger errors were considered harder. These error scores were then converted into sampling weights. Limits were applied to avoid extremely large or small weights and cap any outliers. These weights were passed to a WeightedRandomSampler in PyTorch, which increased the likelihood of selecting harder examples during training. The validation set remained unweighted. The reweighting step was repeated every 10–20 epochs to keep the weights up to date. This method helps the model focus more on difficult configurations while still learning broadly, which leads to better performance on complex regions of the data.

Figure 5(a–c) show parity plots for the M3GNet-MLFF models trained for charge states $q = +1$, $q = 0$, and $q = -1$. Models for the $q = +2$ and $q = -2$ charge states are presented in Figure S8. Each parity plot compares DFT-computed CFE for test set points with the corresponding values from the MLFF prediction for different types of structures: bulk (pristine supercells without defects), defects (bulk supercells containing a single point defect or defect complex), defects at CdTe-ZnTe interfaces, and defects in CdTe grain boundaries. Overall, the MLFF predictions show remarkably low RMSE values for all types of bulk and defect configurations, similar to the ALIGNN models.

For $q = 0$, the test prediction RMSE for bulk structures is 8.6 meV/atom, a similarly low value of 6.8 meV/atom for defect structures, and 3.7 meV/atom for interface. Grain boundary defect structures show slightly larger errors of > 9.1 meV/atom, which is expected given their more complex local environments and atomic rearrangements.

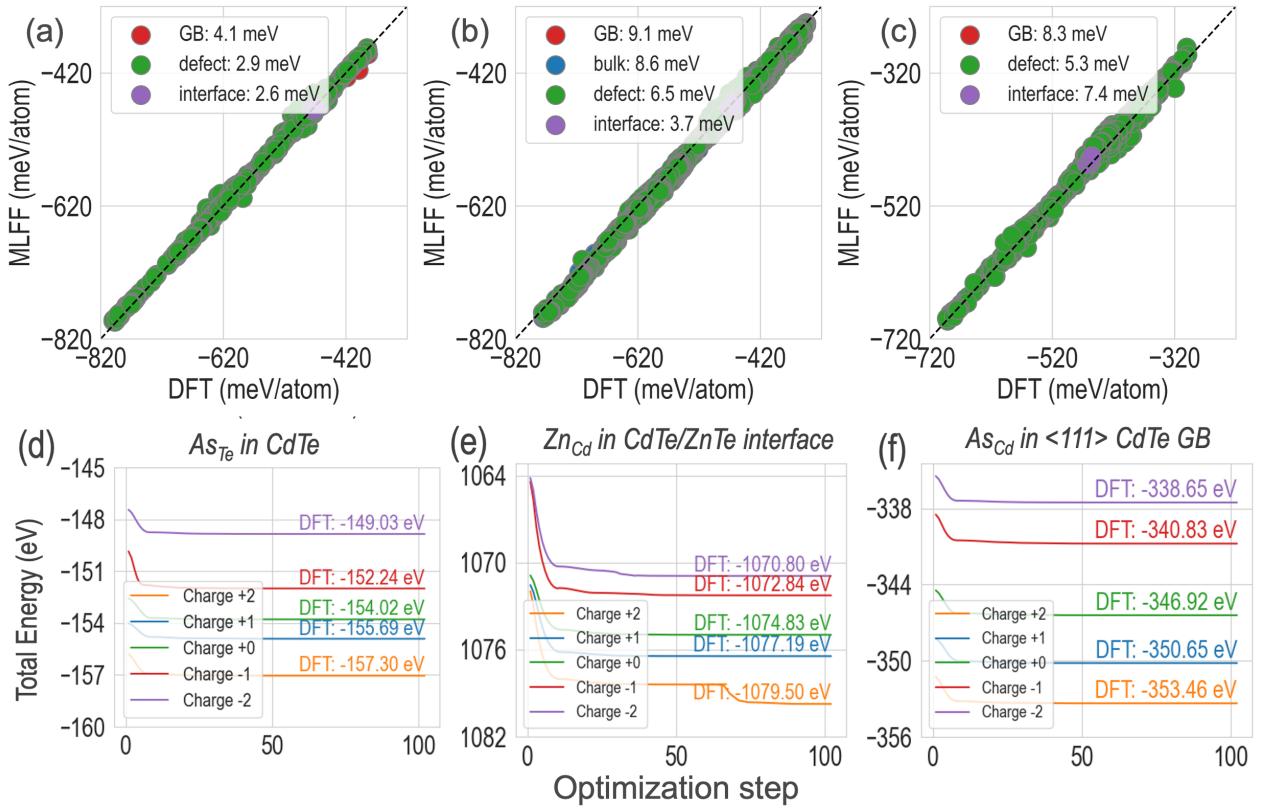


Figure 5 Performance of M3GNet-MLFF models trained on the GGA-PBE dataset, shown in terms of predicted vs DFT crystal formation energy parity plots. The models were trained separately for (a) defect configurations with charge $q=+1$, (b) neutral $q=0$ defect and bulk configurations, and (c) defect configurations with charge $q=-1$. Here, "bulk" refers to pristine supercells without any defects, "defect" means bulk supercells containing a point defect or defect complex, "interface" corresponds to defects located at CdTe-ZnTe interfaces, and "GB" indicates defects situated in CdTe grain boundary structures. Plots in (d–f) show the geometry optimization process taking into account 5 charge states for different example defect configurations: (d) a As_{Te} defect in CdTe , (e) a Zn_{Cd} defect at a CdTe/ZnTe interface, and (f) As_{Cd} in a $<111>$ CdTe grain boundary structure.

Similar errors were seen for charged defect structures as well. The overall agreement between DFT and MLFF predictions remains generally strong, indicating the robustness of the model across diverse chemical and structural environments.

One of the major advantages of having an MLFF model rather than a direct energy prediction model is the ability to use predicted atomic forces to perform geometry optimization based on gradient-based energy minimization, which is more computationally efficient than gradient-free optimization as shown in **Figure S9**. For example, we optimized the $\text{As}_{\text{Te}}+\text{Cl}_{\text{Te}}$ defect in $\text{CdSe}_{0.12}\text{Te}_{0.88}$ using both ALIGNN (direct energy) and M3GNet (MLFF) and found that M3GNet achieved the optimization at a substantially lower computational cost compared to ALIGNN. **Figure 5(d–f)** show three different examples of using the MLFF model for optimizing challenging defect configurations: an As_{Te} substitutional defect in CdTe , a Zn_{Cd} defect in a CdTe-ZnTe interface structure, and an As_{Cd} substitu-

tional defect in a CdTe GB structure. These cases represent chemically and structurally complex environments that are often found in devices with polycrystalline semiconductor thin films. The optimized $q = +2, +1, 0, -1, -2$ MLFF models were able to successfully capture local atomic rearrangements and produce low energy configurations consistent with DFT benchmarks. In each case, the MLFF-optimized configuration energy matches well with the DFT-optimized energy. Energy minimization is achieved in approximately 100 steps, with the entire relaxation process completing within a few minutes.

7 MLFF models trained at hybrid functional accuracy

The ALIGNN and M3GNet-MLFF models trained on the GGA dataset can be enormously useful for optimizing new defect structures and screening for low energy defects. However, because of the well known limitations of the GGA functional in predicting electronic band edges and

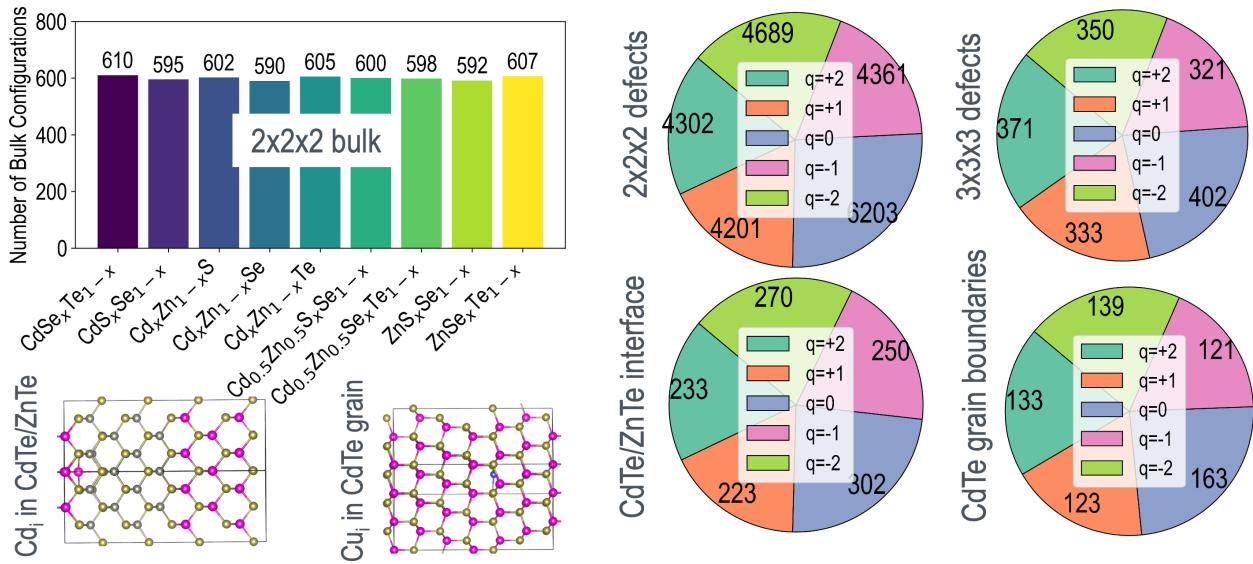


Figure 6 Statistics of the HSE06 dataset: number of bulk configurations corresponding to $\text{CdSe}_x\text{Te}_{1-x}$, $\text{CdS}_x\text{Se}_{1-x}$, $\text{Cd}_x\text{Zn}_{1-x}\text{S}$, $\text{Cd}_x\text{Zn}_{1-x}\text{Se}$, $\text{Cd}_x\text{Zn}_{1-x}\text{Te}$, $\text{Cd}_{0.5}\text{Zn}_{0.5}\text{S}_x\text{Se}_{1-x}$, $\text{Cd}_{0.5}\text{Zn}_{0.5}\text{Se}_x\text{Te}_{1-x}$, $\text{ZnS}_x\text{Se}_{1-x}$, and $\text{ZnSe}_x\text{Te}_{1-x}$ compositions; pie charts showing the number of defect configurations in different charge states in $2 \times 2 \times 2$ supercells, $3 \times 3 \times 3$ supercells, CdTe-ZnTe interfaces, and CdTe GB structures; and, example representations of point defects in interface and grain boundary structures.

Dataset	Supercell Size	Data Points
Bulk dataset from Cd/Zn-Te/Se/S binary compounds and alloys	$2 \times 2 \times 2$	5400
Defect dataset from 6 binary compounds	$2 \times 2 \times 2$	4302 ($q=+2$), 4201 ($q=+1$), 6203 ($q=0$), 4361 ($q=-1$), 4689 ($q=-2$)
Defect dataset from $\text{CdSe}_x\text{Te}_{1-x}$ and $\text{Cd}_x\text{Zn}_{1-x}\text{Te}$ alloys	$3 \times 3 \times 3$	371 ($q=+2$), 333 ($q=+1$), 402 ($q=0$), 321 ($q=-1$), 350 ($q=-2$)
Defect dataset from CdTe/ZnTe interface	$2 \times 2 \times 4$	233 ($q=+2$), 223 ($q=+1$), 302 ($q=0$), 250 ($q=-1$), 270 ($q=-2$)
Defect dataset from CdTe grain boundaries		133 ($q=+2$), 123 ($q=+1$), 163 ($q=0$), 121 ($q=-1$), 139 ($q=-2$)

Table 2 Number of data points (or structures) in the HSE06 dataset corresponding to different types of bulk or defect configurations, supercell sizes, and charge states.

band gaps, chemical potentials, defect energetics, and defect levels⁹, it is important to extend the DFT predictions to the hybrid HSE06 level of theory. Given the significant added computational expense of HSE06 calculations, it will take an extremely long time to generate HSE06 data for the upwards of 20,000 bulk and defect configurations in the GGA-PBE dataset. Thus, we performed HSE06 geometry optimization calculations on a subset of the structures on top of GGA optimization, which includes both initial and newly obtained configurations from the active learning schedule. **Table 2** shows the number of structures of different types (bulk, defects, interfaces, GB) in different charge states and supercell sizes that eventually constituted the HSE06 dataset.

HSE06 calculations for $3 \times 3 \times 3$ supercell defect structures with lattice constants around 20 Å were performed using Γ -point only⁸⁸, 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. Calculations for grain boundary and interface structures were also performed using Γ -point

only. **Figure 6** shows the statistics of the compiled HSE06 dataset in terms of number of bulk Cd/Zn-Te/Se/S composition structures, and different types of defects in different charge states, including interface and GB defects. The HSE06 dataset represents 53.4% of the GGA dataset for bulk ($q = 0$), and 63.8%, 71.6%, 79.5%, 72.4%, and 65.8% for defect 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 CFE values in the HSE06 dataset are presented in **Figure S10**.

Parity plots for M3GNet-MLFF models trained on the HSE06 dataset are pictured in **Figure 7(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 **Figure S11**. Each parity plot compares the HSE06-computed CFE with MLFF-predicted values across different categories: bulk (pristine supercells without defects), defects (bulk supercells containing a single defect or defect complex), defects at CdTe-ZnTe interfaces, and defects

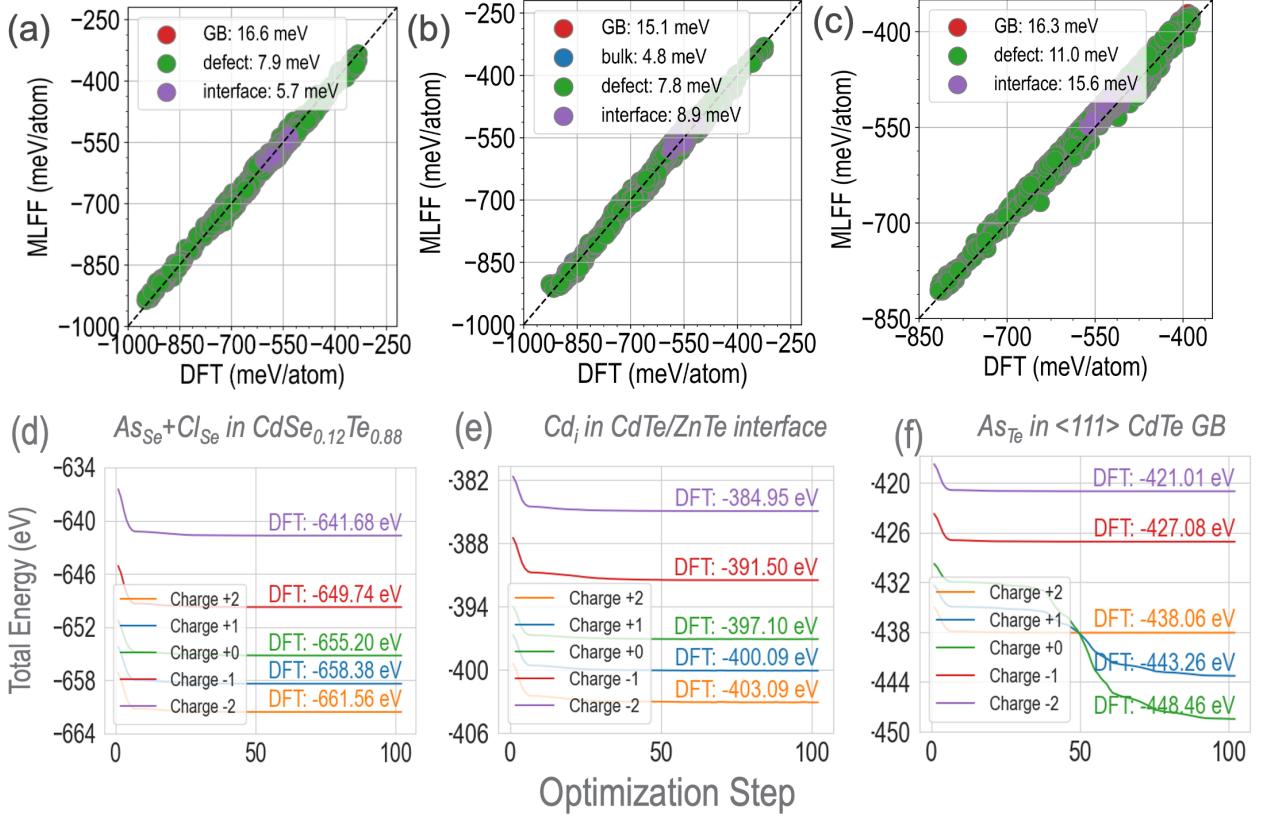


Figure 7 Performance of the M3GNet-MLFF models trained on the augmented HSE06 dataset, shown in terms of predicted vs DFT crystal formation energy parity plots. The models were trained separately for (a) defect configurations with charge $q=+1$, (b) neutral $q=0$ defect and bulk configurations, and (c) defect configurations with charge $q=-1$. Here, "bulk" refers to pristine supercells without any defects, "defect" means bulk supercells containing a point defect or defect complex, "interface" corresponds to defects located at CdTe-ZnTe interfaces, and "GB" indicates defects situated at CdTe grain boundaries. Plots in (d–f) show the geometry optimization process taking into account 5 charge states for different example defect configurations: (d) an $\text{As}_{\text{Se}} + \text{Cl}_{\text{Se}}$ defect complex in $\text{CdSe}_{0.12}\text{Te}_{0.88}$, (e) a Cd_i defect at the CdTe/ZnTe interface, and (f) As_{Te} in $<111>$ CdTe grain boundary structure.

in CdTe GB structures. Despite the reduced dataset 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/atom for bulk structures to close to 7.8 meV/atom for defect structures, 8.9 meV/atom for defects at interfaces, and 15.1 meV/atom for defects in GBs. These errors are slightly larger for $q = +1$ and $q = -1$ defect structures but remain below 20 meV/atom for all cases, which is quite reasonable given the range of CFE values in the dataset. We also simulated selected defects in a grain boundary structure of CdTe and optimized them using the HSE-MLFF model. As illustrated in Figure S12, the MLFF achieved results comparable to full DFT but with a dramatic reduction in computational time: while conventional DFT calculations typically required around 500 minutes, the MLFF completed similar simulations in only a few minutes.

To further evaluate the MLFF's reliability in estimating

defect energetics, we compared its predictions with HSE optimization for both single defects and defect complexes (Figure S13). For neutral defect structures, the MLFF shows excellent correlation with DFT-calculated CFE, with RMSE values of 8.14 meV/atom for single defects and 9.23 meV/atom for complex defects. This suggests that the MLFF effectively captures both localized and collective defect relaxations, even for configurations with multiple defects. Figure 7(d–f) show some examples of using the HSE-MLFF model for optimizing some challenging defect configurations: the $\text{As}_{\text{Se}} + \text{Cl}_{\text{Se}}$ defect complex in $\text{CdSe}_{0.12}\text{Te}_{0.88}$, a Cd_i defect at the CdTe-ZnTe interface, and the As_{Te} substitutional defect in a CdTe GB structure. In each case, the optimization process found the ground state configuration in a few hundred steps, reaching very close to the actual HSE06 computed final energy. This gives us the confidence to use the HSE-MLFF model going forward to efficiently optimize any new defect structures at hybrid functional accuracy, before performing final HSE06+SOC calculations on the most

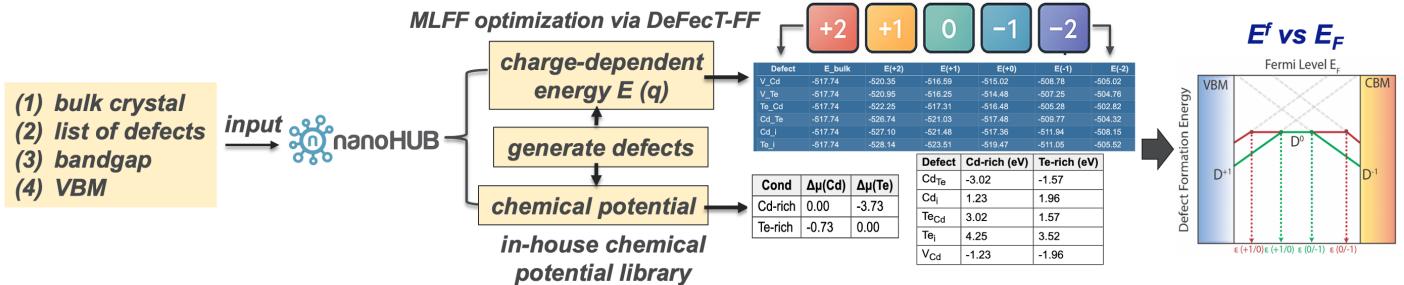


Figure 8 The **DeFecT-FF** tool takes as input the bulk crystal, list of defects, bandgap, and VBM, performs MLFF-based geometry optimization, leverages an in-house chemical potential library to calculate charge-dependent energies $E(q)$ and defect formation energies E_f as a function of the Fermi level E_F , and finally constructs E_f - E_F diagrams for defect thermodynamics. This tool is accessible via a nanoHUB web application.

important defects for validation and discovery. In the next section, we present some of these results, showing how high-fidelity defect formation energy diagrams could be created much faster than performing regular full DFT. Our complete workflow for predicting defect energetics using PBE/HSE-based MLFF models is illustrated in **Figure 4**.

8 HSE-MLFF Case Studies of Important Defects

At this point, the “HSE-MLFF” models can be used to optimize any given defect configurations in different charge states at hybrid functional accuracy. Following this optimization, final HSE06+SOC calculations are necessary for obtaining reliable defect formation energy diagrams. **Figure 8** shows the workflow of a **DeFecT-FF**⁹⁶ 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 sub-sections, we present two case studies where this workflow was applied (using the HSE-MLFF models) to ultimately yield information about the relative stability and transition levels of important defects in chemical compositions of interest.

8.1 As and Cl Defects in CdSe_xTe_{1-x}

Anion-site extrinsic substitutional defects were modeled in $3 \times 3 \times 3$ supercells of multiple CdSe_xTe_{1-x} compositions ($x = 0, 0.06, 0.12, 0.25$). Using the **Doped**⁸⁶ package, we introduced 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. One example is shown in **Figure S14**, where 14 different symmetry-broken configurations were generated for As_{Se}

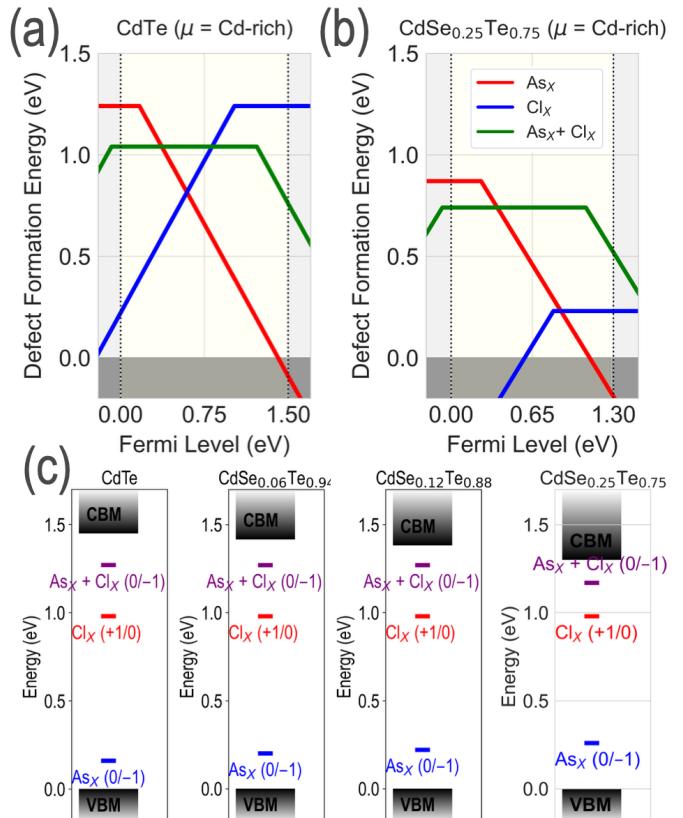


Figure 9 Defect formation energy diagrams for As_X, Cl_X, and As_X+Cl_X defects in (a) CdTe and (b) CdSe_{0.25}Te_{0.75}, under Cd-rich conditions; X = Te or Se. (c) Defect charge transition levels for As_X, Cl_X, and the complex As_X+Cl_X computed for different Se concentrations ($x = 0, 0.06, 0.12, 0.25$) in CdSe_xTe_{1-x}. Blue lines indicate the As_X (0/-1) transition level, red lines show the Cl_X (+1/0) transition, 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 is placed at the band gap. All results are from HSE06+SOC calculations performed after MLFF optimization.

in a CdSeTe composition. The HSE-MLFF model was then applied to identify the lowest energy configuration (configuration 5 in this case). The left panel in **Figure S14** illustrates how standard DFT relaxation may lead to a false ground state, while the right panel shows the relative energies of all generated configurations, with the minimum energy structure highlighted in red.

Hundreds of structures for these substitutional defects across the $\text{CdSe}_x\text{Te}_{1-x}$ compounds were relaxed with the HSE-MLFF models for different charge states until the maximum force fell below $< 10^{-5}$ eV/Å. Finally, single-shot HSE06+SOC calculations were performed on the MLFF-optimized (lowest energy) geometries to obtain accurate defect formation energy, described below for a defect D in a charge state q :

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

Here, $E_{\text{tot}}(D^q)$ and $E_{\text{tot}}(\text{bulk})$ are the total energies of the charged defect supercell and the pristine host supercell respectively, n_i is the number of atoms of species i removed ($n_i > 0$) or added ($n_i < 0$) with chemical potential μ_i , q is the defect charge state, E_{VBM} is the computed valence band maximum energy of the bulk $\text{CdSe}_x\text{Te}_{1-x}$ compound, E_F is the Fermi level which ranges across the band gap, and E_{corr} is the finite-size correction energy⁶. The band gaps computed using HSE06+SOC (with a modified mixing parameter of $\alpha=0.31$) for CdTe, $\text{CdSe}_{0.06}\text{Te}_{0.94}$, $\text{CdSe}_{0.12}\text{Te}_{0.88}$, and $\text{CdSe}_{0.25}\text{Te}_{0.75}$ are respectively 1.5 eV, 1.41 eV, 138 eV, and 1.30 eV; these values are used to place the E_F bounds for the defect formation energy diagrams. 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. **Figure 9** presents the defect formation energy diagrams and relevant transition levels for As_X , Cl_X , and As_X+Cl_X defects across the $\text{CdSe}_x\text{Te}_{1-x}$ 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 $\text{CdSe}_{0.25}\text{Te}_{0.75}$, in agreement with recent experimental studies⁹⁷. 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 **Figure 9(a-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 around the middle of the band gap.

8.2 Native Defects and Nitrogen Impurities in ZnTe

Motivated by experimental evidence from X-ray photoelectron spectroscopy (XPS)⁹⁸⁻¹⁰³ indicating N incorporation in ZnTe^{52,104-109}, we employed the **DeFecT-FF** workflow to systematically investigate both native point defects and N-related defects in ZnTe. A $3 \times 3 \times 3$ ZnTe supercell 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 different N-related defects such as N_i , N_{Te} , and complexes N_i+N_i and $\text{N}_{\text{Te}}+\text{N}_i$. To ensure thorough exploration of the potential energy landscape, we applied ShakeNBreak^{110,111} to induce perturbations, enabling the sampling of a diverse set of competing configurations. Among the hundreds of N-related configurations evaluated, the 2N_i defect emerged as the most energetically favorable, a finding further validated through additional HSE06+SOC calculations. **Figure 10(a)** illustrates the HSE-MLFF structural optimization and energy convergence for the 2N_i defect complex in ZnTe, and **Figure 10(b)** presents the HSE06+SOC computed defect formation energy diagram for N-related defects in ZnTe.

9 Conclusions

$\text{CdSe}_x\text{Te}_{1-x}$ 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 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 introduced the **DeFecT-FF** framework in this work, a crystal graph-based active learning-driven MLFF model trained on both semi-local and hybrid functional calculations for thousands of charged and neutral structures spanning the Cd/Zn-S/Se/Te bulk, alloy, interface, and grain boundary

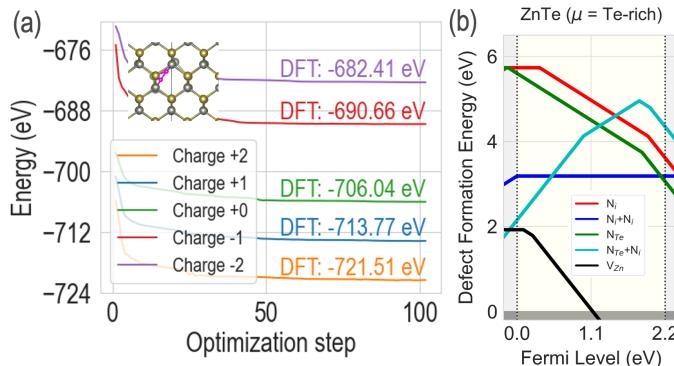


Figure 10 (a) Energy as a function of optimization steps during MLFF relaxation of a ZnTe supercell with the double N interstitial ($2N_i$) defect complex. The inset shows the relaxed configuration with N atoms in red. (b) Defect formation energies in ZnTe under Te-rich conditions computed using HSE06+SOC on top of the HSE-MLFF optimized configurations.

configurations with native and extrinsic defects and defect complexes. 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 \sim 12–14 h (HSE06) to \sim 1–2 min while retaining near-DFT accuracy, transforming comprehensive, composition- and charge-resolved defect surveys from intractable to routine. 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.

Conflicts of Interest

There are no conflicts to declare.

Data Availability

All raw and processed data supporting this work, namely atomic structures, total energies, forces, and stresses, and the trained MLFF models used in this study, are publicly accessible via a nanoHUB web application at <https://nanohub.org/tools/cadetff>. This tool enables community reuse and verification through a point-and-click interface. Within the same tool, users can:

- **Provide inputs:** Upload a crystallographic file (e.g., CIF or POSCAR), select native or extrinsic defect types (vacancy, interstitial, substitutional, or complexes).
- **Run computations:** Perform MLFF-based geometry relaxation in minutes and obtain optimized energies to identify lowest energy configurations.
- **Retrieve outputs:** Download relaxed structures, per-defect energies, and tabulated CSV summaries, along with run logs for reproducibility.
- **Intended scope:** Rapid screening and hypothesis testing; device-critical cases should be validated with targeted HSE06(+SOC) calculations.

Funding Statement

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.

Acknowledgements

The authors would like to acknowledge discussions with Dr. Mariana Bertoni at Arizona State University, Dr. Yanfa Yan at University of Toledo, 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.

Author Contributions Statement

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. M.H.R. took the lead on writing and A.M.-K. contributed in editing and shaping the manuscript.

References

- Wei, *Journal of Applied Physics*, 2016, **119**, 045104.
- 12 D. N. Krasikov, A. V. Scherbinin, A. A. Knizhnik, A. N. Vasiliev, B. V. Potapkin and T. J. Sommerer, *Journal of Applied Physics*, 2016, **119**, 085706.
- 13 D. Krasikov, A. Knizhnik, B. Potapkin, S. Selezneva and T. Sommerer, *Thin Solid Films*, 2013, **535**, 322–325.
- 14 D. Krasikov, D. Guo, S. Demtsu and I. Sankin, *Solar Energy Materials and Solar Cells*, 2021, **224**, 111012.
- 15 D. Krasikov and I. Sankin, *Physical Review Materials*, 2018, **2**, 103803.
- 16 D. Krasikov, *Nat. Energy*, 2019, **4**, 442–443.
- 17 T. Ablekim, S. K. Swain, W.-J. Yin, K. Zaunbrecher, J. Burst, T. M. Barnes, D. Kuciauskas, S.-H. Wei and K. G. Lynn, *Scientific Reports*, 2017, **7**, 4563.
- 18 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.
- 19 P. Gorai, D. Krasikov, S. Grover, G. Xiong, W. K. Metzger and V. Stevanović, *Science Advances*, 2023, **9**, eade3761.
- 20 R. De Souza and G. Harrington, *Nature Materials*, 2023, **22**, 794–797.
- 21 D. V. Lang, *Journal of Applied Physics*, 2003, **45**, 3023–3032.
- 22 D. Wickramaratne, C. E. Dreyer, B. Monserrat, J.-X. Shen, J. L. Lyons, A. Alkauskas and C. G. Van de Walle, *Applied Physics Letters*, 2018, **113**, 192106.
- 23 J. Y. Kim, L. Gelczuk, M. P. Polak, D. Hlushchenko, D. Morgan, R. Kudrawiec and I. Szlufarska, *Npj 2D Materials and Applications*, 2022, **6**, 75.
- 24 D. Broberg, K. Bystrom and et al, *npj Computational Materials*, 2023, **9**, 72.
- 25 M. Y. Toriyama, J. Qu, G. J. Snyder and P. Gorai, *J. Mater. Chem. A*, 2021, **9**, 20685–20694.
- 26 C.-W. Lee, N. U. Din, K. Yazawa, W. Nemeth, R. W. Smaha, N. M. Haegel and P. Gorai, *Journal of Applied Physics*, 2024, **135**, 155101.
- 27 R. Grill and A. Zappettini, *Prog. Cryst. Growth Charact. Mater.*, 2004, **48-49**, 209–244.
- 28 J. Buckeridge, *Computer Physics Communications*, 2019, **244**, 329–342.
- 29 A. Mannodi-Kanakkithodi, X. Xiang, L. Jacoby, R. Biegaj, S. T. Dunham, D. R. Gamelin and M. K. Y. Chan, *Patterns (N. Y.)*, 2022, **3**, 100450.
- 30 M. H. Rahman and A. Mannodi-Kanakkithodi, *Defect Generation in Graph Neural Networks*, 2023, https://github.com/msehabibur/defect_

- GNN_gen_1, GitHub repository.
- 31 A. Mannodi-Kanakkithodi, J.-S. Park, A. B. F. Martinson and M. K. Y. Chan, *The Journal of Physical Chemistry C*, 2020, **124**, 16729–16738.
 - 32 S. Kim, J.-S. Park, S. Hood and A. Walsh, *Journal of Materials Chemistry A*, 2019, **7**, 2686–2693.
 - 33 J. L. Lyons and C. G. Van de Walle, *Npj Comput. Mater.*, 2017, **3**, 1–10.
 - 34 A. Machín and F. Márquez, *Materials*, 2024, **17**, 1165.
 - 35 I. Mosquera-Lois, J. Klarbring and A. Walsh, *Chemical Science*, 2025.
 - 36 I. Mosquera-Lois, S. R. Kavanagh, J. Klarbring, K. Tølborg and A. Walsh, *Chemical Society Reviews*, 2023, **52**, 5812–5826.
 - 37 M. P. Polak, R. Jacobs, A. Mannodi-Kanakkithodi, M. K. Y. Chan and D. Morgan, *The Journal of Chemical Physics*, 2022, **156**, 114110.
 - 38 A. Mannodi-Kanakkithodi and M. K. Y. Chan, *Journal of Materials Science*, 2022, **57**, 10736–10754.
 - 39 T. Xie and J. C. Grossman, *Phys. Rev. Lett.*, 2018, **120**, 145301.
 - 40 K. Choudhary and B. G. Sumpter, *AIP Advances*, 2023, **13**, 095109.
 - 41 Z. Chen, X. Li and J. Bruna, *Supervised Community Detection with Line Graph Neural Networks*, 2017, <https://arxiv.org/abs/1705.08415v6>.
 - 42 T. N. Kipf and M. Welling, *Semi-Supervised Classification with Graph Convolutional Networks*, 2016, <https://arxiv.org/abs/1609.02907v4>.
 - 43 M. D. Witman, A. Goyal, T. Ogitsu, A. H. McDaniel and S. Lany, *Nature Computational Science*, 2023.
 - 44 C. Chen and S. P. Ong, *Nature Computational Science*, 2022, **2**, 718–728.
 - 45 M. H. Rahman, P. Gollapalli, P. Manganaris, S. K. Yadav, G. Pilania, B. DeCost, K. Choudhary and A. Mannodi-Kanakkithodi, *APL Machine Learning*, 2024, **2**, 016122.
 - 46 M. H. Rahman, S. Rojsatien, D. Krasikov, M. K. Chan, M. Bertoni and A. Mannodi-Kanakkithodi, *Solar Energy Materials and Solar Cells*, 2025, **293**, 113857.
 - 47 M. H. Rahman and A. Mannodi-Kanakkithodi, *Journal of Physics Materials*, 2025, **8**, 022001.
 - 48 J. Pan, W. K. Metzger and S. Lany, *Phys. Rev. B*, 2018, **98**, 054108.
 - 49 P. Borlido, J. Schmidt, A. W. Huran, F. Tran, M. A. L. Marques and S. Botti, *npj Computational Materials*, 2020, **6**.
 - 50 C. Vona, D. Nabok and C. Draxl, *Advanced Theory and Simulations*, 2022, **5**, 2100496.
 - 51 J. L. Lyons and C. G. Van de Walle, *npj Computational Materials*, 2017, **3**, 12.
 - 52 E. Menéndez-Proupin, M. Casanova-Páez, A. L. Montero-Alejo, M. A. Flores and W. Orellana, *Physica B Condensed Matter*, 2019, **568**, 81–87.
 - 53 F. K. Alfadhili, A. B. Phillips, G. K. Liyanage, J. M. Gibbs, M. K. Jamarkattel and M. J. Heben, *MRS Advances*, 2019, **4**, 913–919.
 - 54 O. de Melo, M. Behar, J. F. Dias, R. Ribeiro-Andrade, M. da Silva, A. G. de Oliveira and J. C. González, *Materials Science in Semiconductor Processing*, 2019, **97**, 17–20.
 - 55 K. Luo, W. Wu, S. Xie, Y. Jiang, S. Liao and D. Qin, *Applied Sciences*, 2019, **9**, 1885–1885.
 - 56 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 Cu₂ZnSn(S,Se)4 solar cell efficiency*, 2019, <https://www.sciencedirect.com/science/article/pii/S2468606919300097>.
 - 57 K. Shen, X. Wang, Y. Zhang, H. Zhu, Z. Chen, C. Huang and Y. Mai, *Solar Energy*, 2020, **201**, 55–62.
 - 58 J. Miao, X. Liu, K. Jo, K. He, R. Saxena, B. Song, H. Zhang, J. He, M. Han, W. Hu and D. Jariwala, *Nano Letters*, 2020, **20**, 2907–2915.
 - 59 A. G. García and S. Zarate, *Microscopy and Microanalysis*, 2020, **26**, 2804–2805.
 - 60 X. Zheng, E. Colegrove, J. N. Duenow, J. Moseley and W. K. Metzger, *Journal of Applied Physics*, 2020, **128**, 053102.
 - 61 A. Wardak, W. Chromiński, A. Reszka, D. Kochanowska, M. Witkowska-Baran, M. Lewandowska and A. Mycielski, *Journal of Alloys and Compounds*, 2021, **874**, 159941–159941.
 - 62 P. D. Hatton, M. J. Watts, Y. Zhou, R. Smith and P. Goddard, *Journal of Physics Condensed Matter*, 2022, **35**, 75702–75702.
 - 63 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, *Solar Energy Materials and Solar Cells*, 2023, **255**, 112289–112289.
 - 64 Y. Huang, S. R. Kavanagh, D. O. Scanlon, A. Walsh and R. L. Z. Hoye, *Nanotechnology*, 2020, **32**, 132004–132004.
 - 65 T. Bidaud, J. Moseley, M. Amarasinghe, M. Al-Jassim,

- W. K. Metzger and S. Collin, *Imaging CdCl₂ defect passivation and formation in polycrystalline CdTe films by cathodoluminescence*, 2021, <https://journals.aps.org/prmaterials/abstract/10.1103/PhysRevMaterials.5.064601>.
- 66 J. Shi and M. Zikry, *Materials Science and Engineering A*, 2009, **520**, 121–133.
- 67 Y. Zhao, E. M. D. Siriwardane, Z. Wu, N. Fu, M. Al-Fahdi, M. Hu and J. Hu, *npj Computational Materials*, 2023, **9**, 38.
- 68 S. Manna, H. Chan, A. Ghosh, T. Chakrabarti and S. K. R. S. Sankaranarayanan, *Computational Materials Science*, 2023, **229**, 112384–112384.
- 69 M. Cheng, C.-L. Fu, B. Yu, E. Rha, A. Chotrattanapituk, D. L. Abernathy, Y. Cheng and M. Li, 2025.
- 70 M. Jin, J. Miao, M. Khafizov, B. Chen, Y. Zhang and D. H. Hurley, 2025.
- 71 D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue and T. Lookman, *Nature Communications*, 2016, **7**, 11241.
- 72 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, *Scientific Reports*, 2014, **4**, 6367.
- 73 F. Ren, L. Ward, T. Williams, K. J. Laws, C. Wolverton, J. Hattrick-Simpers and A. Mehta, *Science Advances*, 2018, **4**, eaalq1566.
- 74 C. Kim, A. Chandrasekaran, A. Jha and R. Ramprasad, *MRS Communications*, 2019, **9**, 860–866.
- 75 J. C. Verduzco, E. E. Marinero and A. Strachan, *Integrating Materials and Manufacturing Innovation*, 2021, **10**, 299–310.
- 76 J. E. Gentle, *Computational Statistics*, 2010, <https://doi.org/10.1016/b978-0-08-044894-7.01316-6>.
- 77 D. E. Farache, J. C. Verduzco, Z. D. McClure, S. Desai and A. Strachan, *Computational Materials Science*, 2022, **209**, 111386.
- 78 J. Heyd, G. E. Scuseria and M. Ernzerhof, *The Journal of Chemical Physics*, 2003, **118**, 8207–8215.
- 79 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, *Nature Physics*, 2020, **16**, 448–454.
- 80 C. Li, J. Poplawsky, Y. Yan and S. J. Pennycook, *Mater. Sci. Semicond. Process.*, 2017, **65**, 64–76.
- 81 C. Battaglia, A. Cuevas and S. De Wolf, *Energy Environ. Sci.*, 2016, **9**, 1552–1576.
- 82 A. Mannodi-Kanakkithodi, *Modelling and Simulation in Materials Science and Engineering*, 2022, **30**, 044001.
- 83 A. Mannodi-Kanakkithodi, M. Y. Toriyama, F. G. Sen, M. J. Davis, R. F. Klie and M. K. Y. Chan, *npj Computational Materials*, 2020, **6**, 39.
- 84 A. Zunger, S.-h. Wei, L. G. Ferreira and J. E. Bernard, *Physical Review Letters*, 1990, **65**, 353–356.
- 85 F. G. Sen, A. Mannodi-Kanakkithodi, T. Paulauskas, J. Guo, L. Wang, A. Rockett, M. J. Kim, R. F. Klie and M. K. Chan, *Solar Energy Materials and Solar Cells*, 2021, **232**, 111279.
- 86 S. R. Kavanagh, A. G. Squires, A. Nicolson, I. Mosquera-Lois, A. M. Ganose, B. Zhu, K. Brlec, A. Walsh and D. O. Scanlon, *The Journal of Open Source Software*, 2024, **9**, 6433.
- 87 K. Choudhary and B. DeCost, *npj Computational Materials*, 2022, **8**, 221.
- 88 I. Mosquera-Lois, S. R. Kavanagh, A. M. Ganose and A. Walsh, *Npj Computational Materials*, 2024, **10**, 121.
- 89 K. Choudhary, B. DeCost, L. Major, K. Butler, J. Thiya-galingam and F. Tavazza, *Digital Discovery*, 2023, **2**, 346–355.
- 90 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chemistry of Materials*, 2019, **31**, 3564–3572.
- 91 J. Cheng, C. Zhang and L. Dong, *Communications Materials*, 2021, **2**, 92.
- 92 J. Lee and R. Asahi, *Computational Materials Science*, 2021, **190**, 110314.
- 93 V. Fung, J. Zhang, E. Juarez and B. G. Sumpter, *npj Computational Materials*, 2021, **7**, 84.
- 94 S. R. Bahn and K. W. Jacobsen, *Comput. Sci. Eng.*, 2002, **4**, 56–66.
- 95 A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Petersen, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng and K. W. Jacobsen, *Journal of Physics: Condensed Matter*, 2017, **29**, 273002.
- 96 M. H. Rahman and A. K. M. Kanakkithodi, *Introducing DeFecT-FF for Accelerated Modeling of Defect Thermodynamics in CdSeTe Solar Cells*, 2025, <https://nanohub.org/resources/cadetff>.
- 97 P. Ščajev, M. Nardone, C. Reich, R. Farshchi, K. McReynolds, D. Krasikov and D. Kuciauskas, *Advanced Energy Materials*, 2024, 2403902.
- 98 F. A. Stevie and C. L. Donley, *Journal of Vacuum Sci-*

- ence & Technology A Vacuum Surfaces and Films, 2020, **38**, 063204.
- 99 J. Mahoney, C. A. Monroe, A. M. Swartley, M. G. Ucak-Astarlioglu and C. A. Zoto, *Spectroscopy Letters*, 2020, **53**, 726–736.
- 100 A. Born, F. O. L. Johansson, T. Leitner, D. Kühn, A. Lindblad, N. Mårtensson and A. Föhlisch, *Scientific Reports*, 2021, **11**, 16596.
- 101 G. Lanza, M. J. Jimenez, F. Alvarez, J. Pérez and A. Ávila, *ACS Omega*, 2022, **7**, 34521–34527.
- 102 H. Chen, D. T. L. Alexander and C. Hébert, *Nano Letters*, 2024, **24**, 10177–10185.
- 103 H. Xie, X. Cheng and H. Huang, *Investigation on the Interfaces in Organic Devices by Photoemission Spectroscopy*, 2025, <https://doi.org/10.3390/nano15090680>.
- 104 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, *Chemical Communications*, 2015, **51**, 7463–7465.
- 105 L. Zhao, C. Sun, G. Tian and Q. Pang, *Journal of Colloid and Interface Science*, 2017, **502**, 1–7.
- 106 Y. Li, G. Zha, D. Wei, F. Yang, J. Dong, S. Xi, L. Xu and W. Jie, *Sensors*, 2020, **20**, 2032–2032.
- 107 T. Li, Y. Zhu, X. Ji, W. Zheng, Z. Lin, X. Lu and F. Huang, *The Journal of Physical Chemistry Letters*, 2020, **11**, 8901–8907.
- 108 D. Dragoni, T. D. Daff, G. Csányi and N. Marzari, *Physical Review Materials*, 2018, **2**, 013808.
- 109 E. Berger, M. Bagheri and H. Komsa, *Small*, 2025.
- 110 I. Mosquera-Lois, S. R. Kavanagh, A. Walsh and D. O. Scanlon, *Npj Computational Materials*, 2023, **9**, 25.
- 111 I. Mosquera-Lois, S. R. Kavanagh, A. Walsh and D. O. Scanlon, *The Journal of Open Source Software*, 2022, **7**, 4817.
- 112 G. Cheng, X.-G. Gong and W.-J. Yin, *Nature Communications*, 2022, **13**, 1492.

Supplemental material to "DeFecT-FF: Accelerated Modeling of Defects in Cd/Zn-Te/Se/S Compounds combining High-Throughput DFT and Machine Learning Force Fields"

Md Habibur Rahman,¹⁾ and Arun Mannodi-Kanakkithodi ^{1, a)}

¹School of Materials Engineering, Purdue University, West Lafayette, Indiana 47907, USA

Active Learning (AL) Workflow

- A. **Training the Ensemble of ALIGNN Models:** We begin by training an ensemble of ALIGNN models to capture the variability and uncertainties associated with the predictions. To make this ensemble, we partition the original training set into multiple subsets, each containing a different combination of training, validation, and test data. A total of 100 different ALIGNN models have been trained, each on a unique subset of the data, allowing us to account for variability due to data partitioning. **Figure 3(b)** illustrates the ALIGNN predictions on the test dataset (we name it ALIGNN-1) vs. DFT calculations from 100 different ALIGNN models, highlighting the standard deviation in the predictions along with the mean.
- B. **Prediction Across Expanded Defect Chemical Space:** After training the ensemble of ALIGNN-1 models, we utilize them to predict the CFE of all the defects in the expanded chemical space. For each configuration, predictions are made using all 100 models in the ensemble, yielding a distribution of predictions. This approach enables us to not only obtain the mean prediction but also to quantify the uncertainty associated with each prediction. **Figure S4(a)** shows the violin plot of predicted mean CFE (averaged across 100 ALIGNN models) made across the entire set of defects.
- C. **Uncertainty Quantification:** The uncertainty of each prediction is quantified by analyzing the standard deviation of CFE among the predictions made by the 100 ALIGNN-1 models. In our AL framework, we employed the maximum uncertainty (MU) acquisition function⁷⁷. The MU criterion is defined as $MU(x) = \sigma(x)$, where $\sigma(x)$ denotes the standard deviation (uncertainty) of the prediction. **Figure S4(b)** highlights the defects that maximize the MU acquisition function identified by the ALIGNN-1 models.
- D. **Active Learning via Bayesian optimization and New DFT Calculations:** We utilized Bayesian optimization to refine the predictions of the ALIGNN-1 models by selecting the 200 configurations that maximize the chosen acquisition function, prioritizing the most informative data points. These selected configurations were then used to launch new DFT calculations, ensuring that the model iteratively improves its accuracy and predictive performance. Although the model initially has not encountered certain defects, such as those in the $Cd_xZn_{1-x}Te$ composition, the predictions from ALIGNN-1 models are reasonable. **Figure S4(c)** shows ALIGNN-1 prediction (energy of the initial input defect structure, *viz.*, unoptimized energy) vs. DFT calculation (unoptimized energy) for 200 selected defects based on acquisition function. To further improve the performance of ALIGNN-1 models, these new calculations are then incorporated into the training set, iteratively improving the accuracy of the model.
- E. **Final Model Performance:** Following the first iteration of the AL loop, we retrain ALIGNN-1 models and get new ALIGNN models (we name it ALIGNN-2) which are used to make predictions across the remaining defect space. The violin plot of predicted mean CFE (averaged across 100 ALIGNN-2 models) made on the remaining defect configurations is presented in **Figure S4(d)**. The predictions are again evaluated using the acquisition function as shown in **Figure S4(e)**, and 200 new configurations that maximize its value are selected for new DFT calculations. We observe a significant improvement in the ALIGNN predictions as depicted in **Figure S4(f)**, closely matching the DFT-computed CFE across the selected defects. This rapid convergence after just one training cycle highlights the effectiveness of the AL approach in enhancing model performance, even in underexplored regions of the chemical space. Later on,

^aamannodi@purdue.edu

newly obtained DFT data is again incorporated into the training set and we retrained the model (ALIGNN-3). Finally, the ALIGNN-3 models are used to predict the remaining unexplored defects. The rationale behind choosing the 200 defects that maximizes the MU acquisition function is driven by a careful consideration of our computational budget and capabilities. While we could have selected more or fewer defects, choosing 200 (roughly 1.5% of the entire expanded space of defects) represents an optimal balance between maximizing information gain and managing our computational resources effectively.

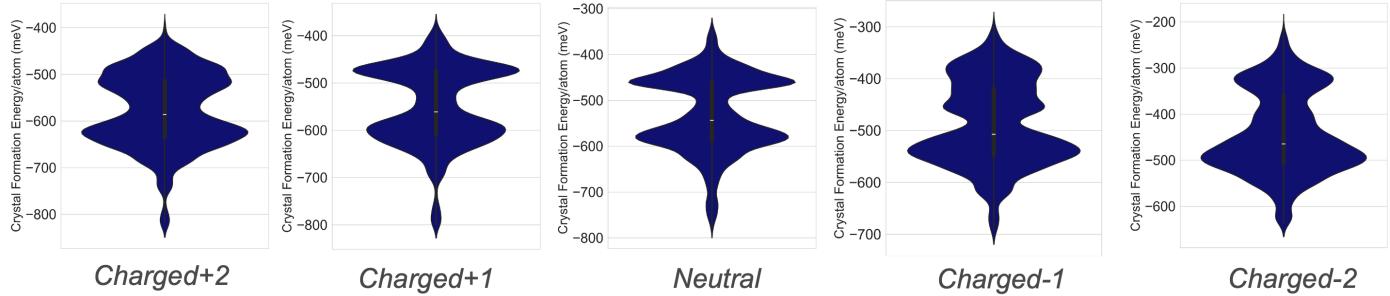


Figure S1 Violin distributions of crystal formation energy per atom (meV) for charge states +2, +1, 0 (neutral), -1, and -2 in the GGA-PBE dataset.

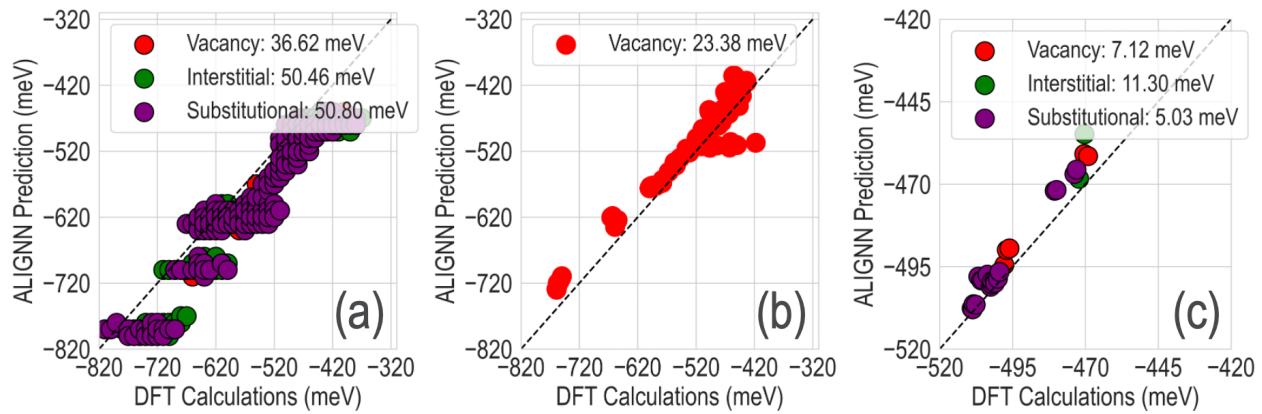


Figure S2 Parity plots comparing ALIGNN predictions to DFT calculations under different training conditions: (a) ALIGNN (trained solely on bulk data) predictions vs. DFT calculations for the defect dataset. (b) ALIGNN (trained exclusively on interstitial defect data) predictions vs. DFT calculations for the vacancy dataset. (c) ALIGNN (trained on a $2 \times 2 \times 2$ supercell bulk+defect data) predictions vs. DFT calculations for defects in a $3 \times 3 \times 3$ supercell.

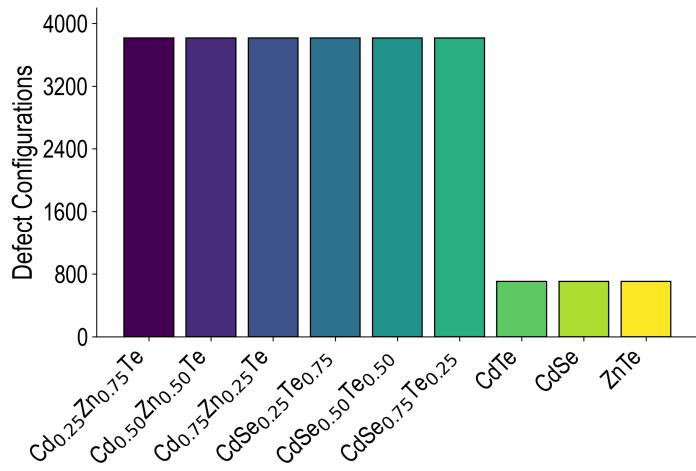


Figure S3 Distribution of extended defect space across the 9 Cd-Se-Te and Cd-Zn-Te compounds.

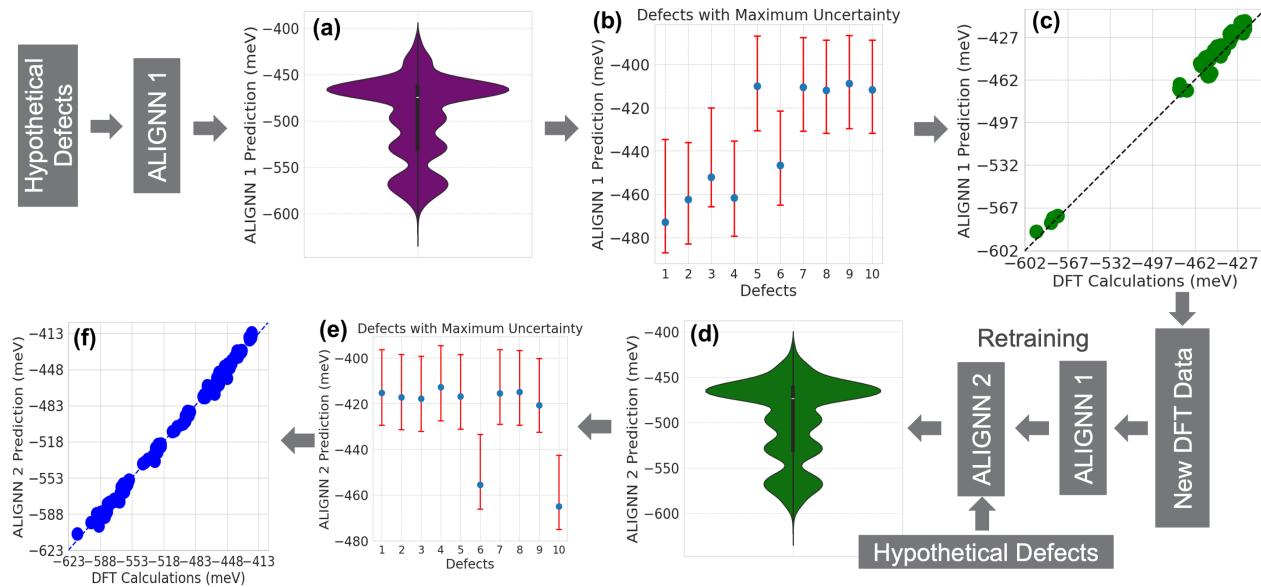


Figure S4 (a) Violin plot showing the mean crystal formation energy (CFE) predicted by the initial ALIGNNN-1 models, averaged across 100 models. (b) Defects with maximum uncertainties identified through ALIGNNN-1 models. (c) Comparison of ALIGNNN-1 predictions vs DFT calculations for 200 selected defects. (d) Violin plot of mean CFE from ALIGNNN-2 models. (e) Defects with maximum uncertainties identified through ALIGNNN-2 models. (f) Comparison of ALIGNNN 2 predictions vs DFT calculations for the 200 selected defects.

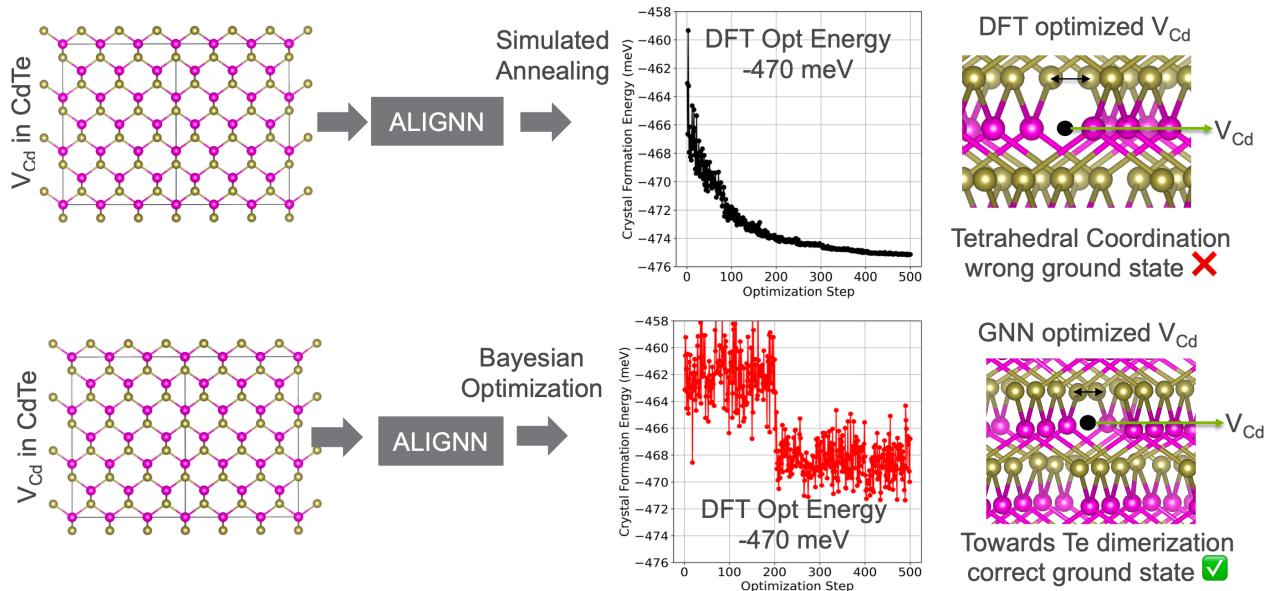


Figure S5 Optimization of Cd vacancy (V_{Cd}) in CdTe using the trained ALIGNNN model with two different optimization strategies: simulated annealing and Bayesian optimization.

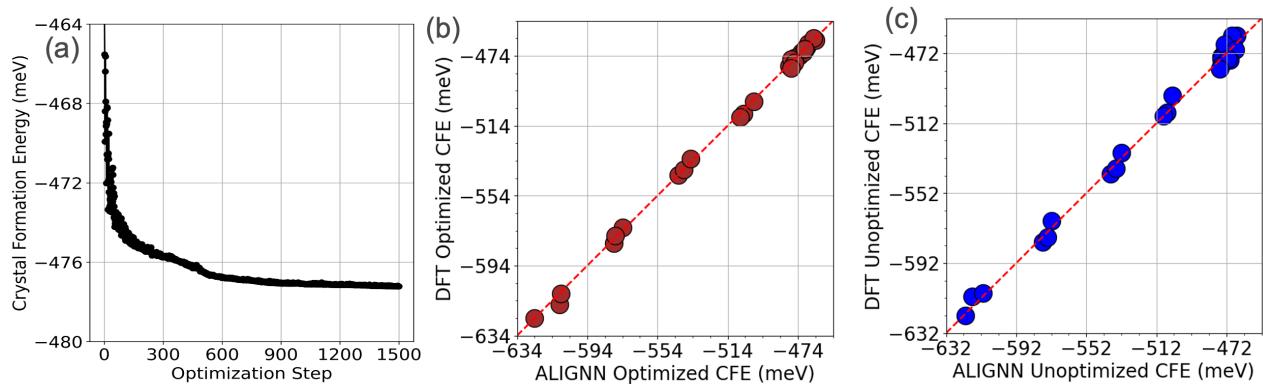


Figure S6 a) Simulated annealing optimization of a Cl_{Te} defect in $\text{CdSe}_{0.50}\text{Te}_{0.50}$ using ALIGNNN-predicted crystal formation energy (CFE), showing energy minimization over successive steps. (b) Parity plot comparing ALIGNNN-optimized CFE with DFT-optimized CFE across a set of defect structures, demonstrating strong agreement. (c) Parity plot comparing ALIGNNN-unoptimized CFE with DFT-unoptimized CFE.

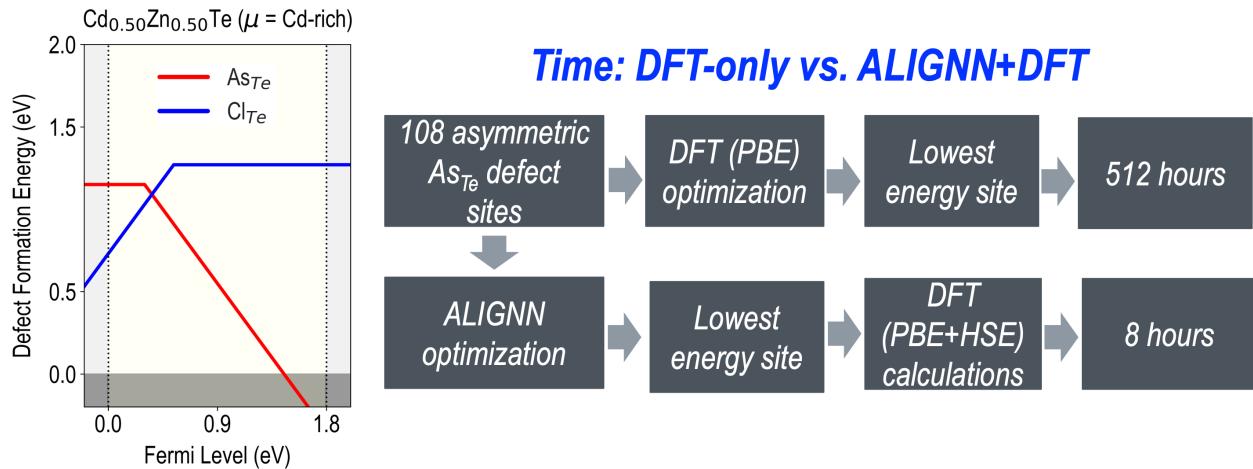


Figure S7 Comparison of computational efficiency between a DFT-only workflow and the ALIGNN+SA+DFT approach for evaluating Cl_{Te} and As_{Te} defects in Cd_{0.50}Zn_{0.50}Te under Cd-rich conditions. Left: Defect formation energies (DFEs) as a function of Fermi level, computed using the HSE06+SOC functional on top of the PBE optimized structures. Right: Workflow comparison showing that direct DFT relaxation of all 108 symmetry-inequivalent As_{Te} configurations requires approximately 512 hours, while the ALIGNN+SA model identifies the lowest-energy site in minutes, followed by ALIGNN+PBE optimization. A single static HSE06+SOC calculation on the ALIGNN+PBE optimized configuration reduces the total computational time to just 8 hours.

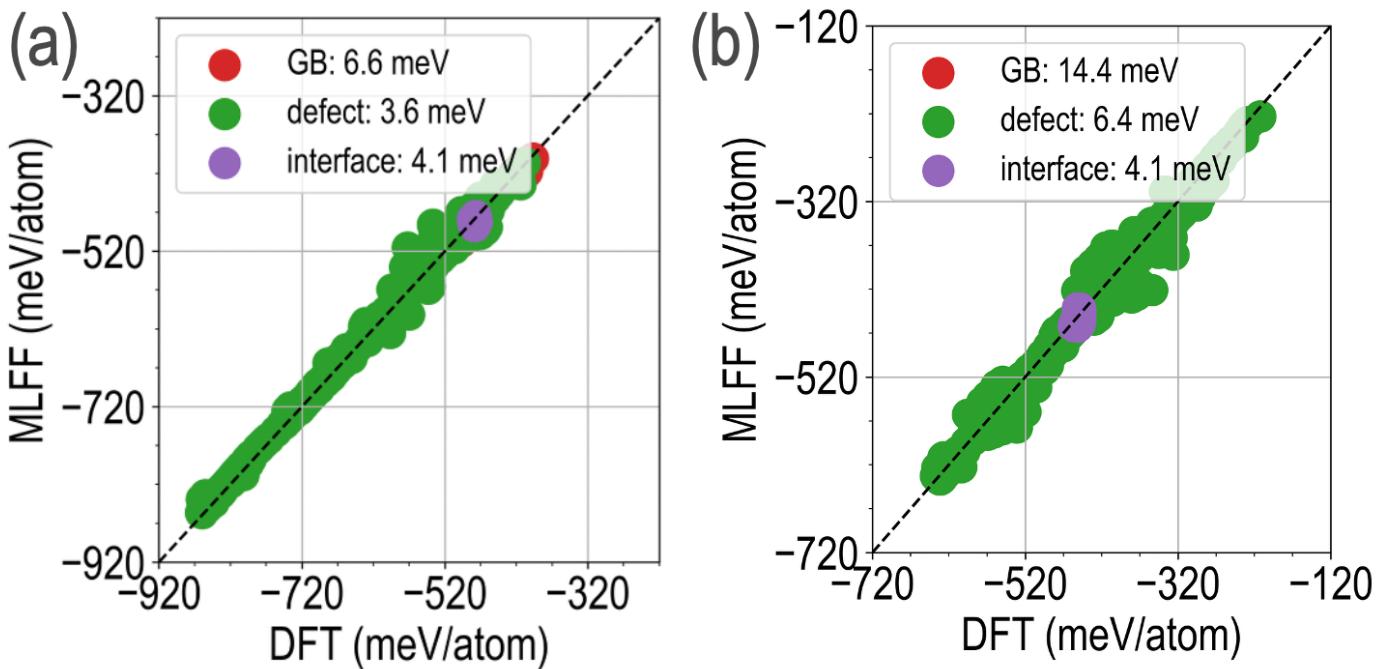


Figure S8 Parity plots for M3GNet-MLFF models trained on the GGA dataset, shown in terms of predicted vs actual (from DFT) crystal formation energies, trained separately for (a) defect configurations with charge $q=+2$, (b) defect configurations with charge $q=-2$. Here, "defect" represents bulk supercells containing a point defect or defect complex, "interface" corresponds to defects located at CdTe-ZnTe interfaces, and "GB" indicates defects situated at CdTe grain boundaries.

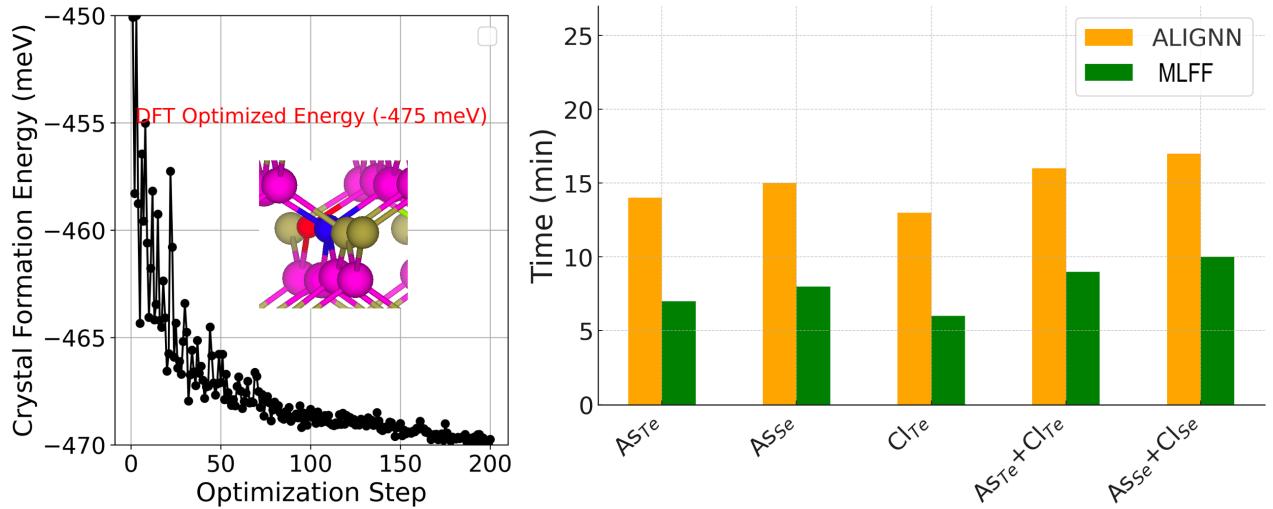


Figure S9 (a) Optimization of an $\text{As}_{Te} + \text{Cl}_{Te}$ complex in $\text{CdSe}_{0.12}\text{Te}_{0.88}$ using an M3GNET-MLFF model, and (b) comparison of the time taken by ALIGNN and the MLFF for optimizing selected defects in $\text{CdSe}_{0.12}\text{Te}_{0.88}$.

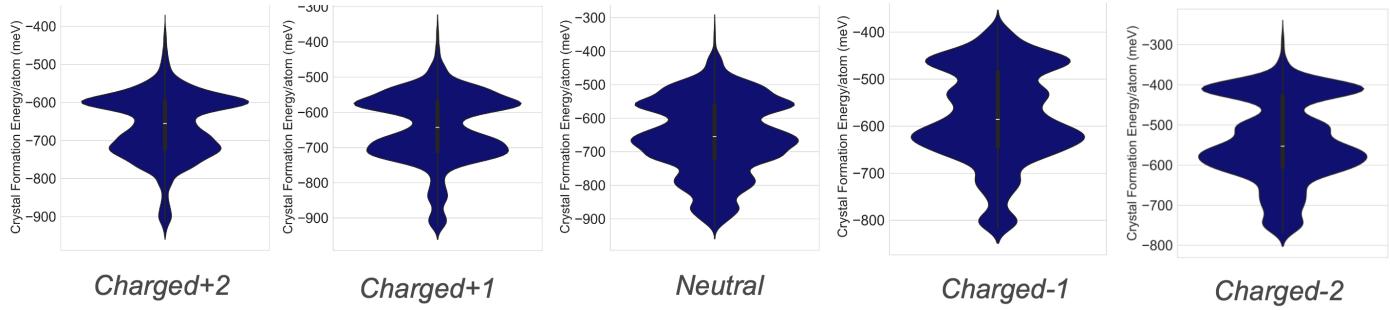


Figure S10 Violin distributions of crystal formation energy per atom (meV) for charge states +2, +1, 0 (neutral), -1, and -2 in the HSE06 dataset.

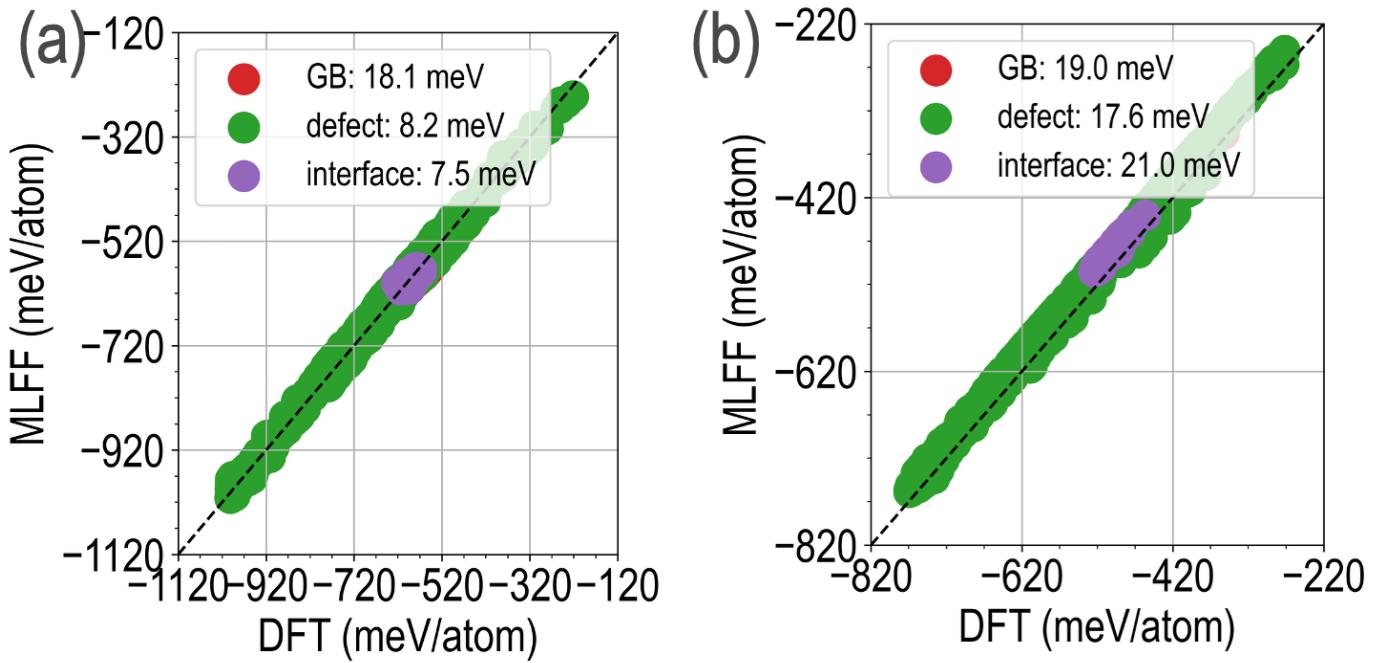


Figure S11 Parity plots for M3GNet-MLFF models trained on the HSE06 dataset, shown in terms of predicted vs actual (from DFT) crystal formation energies, trained separately for (a) defect configurations with charge $q=+2$, (b) defect configurations with charge $q=-2$. Here, "defect" represents bulk supercells containing a point defect or defect complex, "interface" corresponds to defects located at CdTe-ZnTe interfaces, and "GB" indicates defects situated at CdTe grain boundaries.

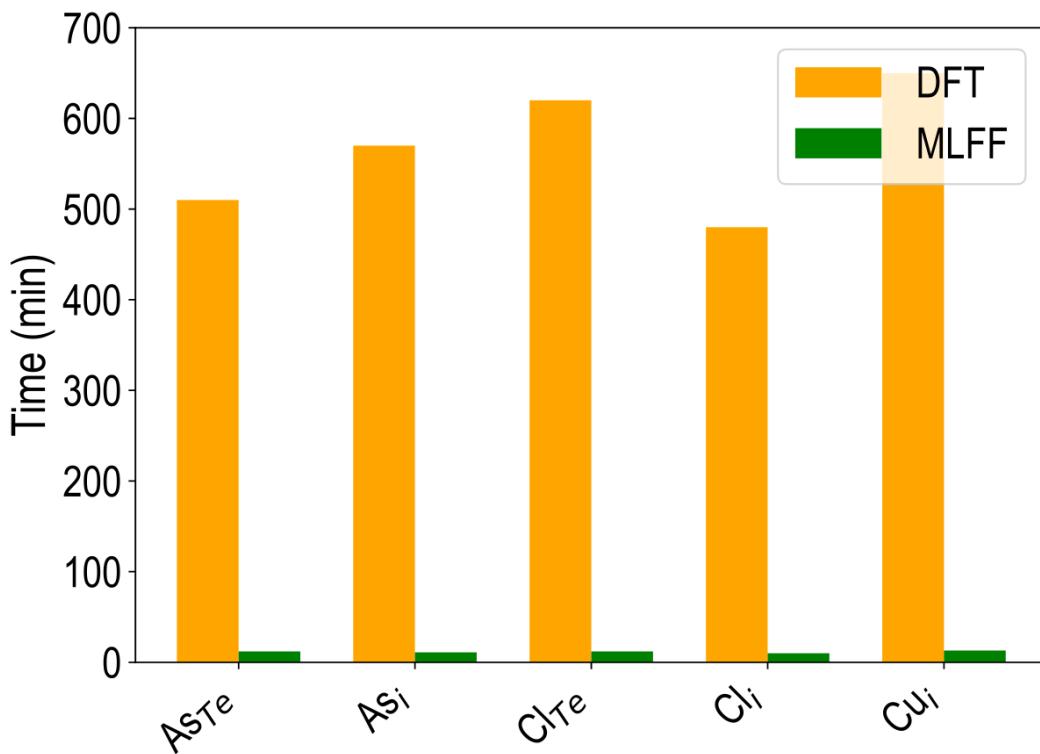


Figure S12 Comparison of computational time between DFT and MLFF for simulating selected defects at the CdTe grain boundary. While DFT calculations typically required around 500–650 minutes, the MLFF achieved similar accuracy within only a fraction of a minute, demonstrating a significant acceleration in simulation speed.

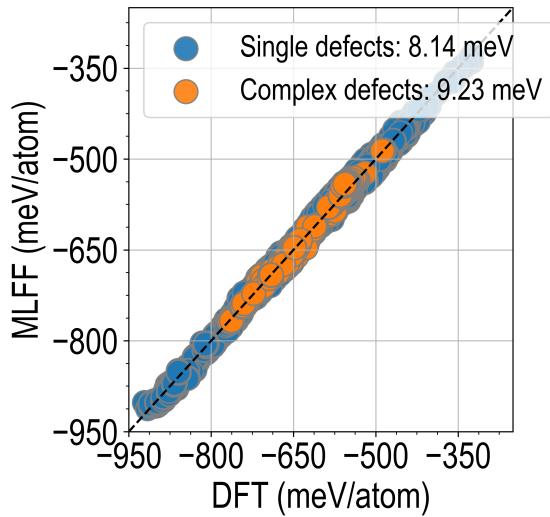
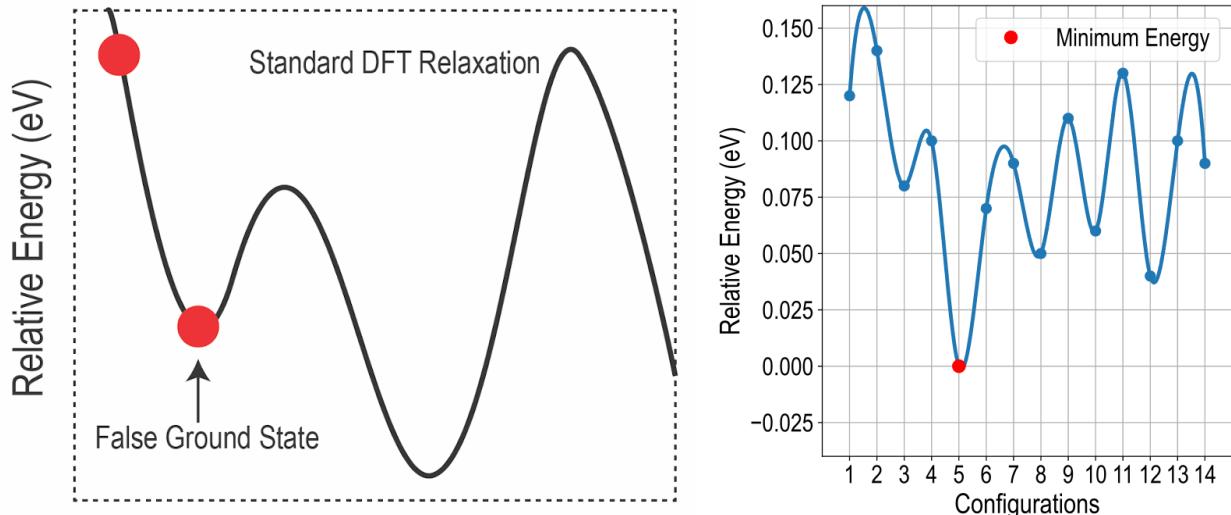


Figure S13 Parity plot comparing DFT and MLFF-predicted crystal formation energies (CFE) for single and complex defects ($q=0$ charge state).



Details on GNN-Based Geometry Optimization

In this study, we employed the ALIGNN model, trained on trajectories obtained from PBE calculations, for optimizing crystal structures containing point defects using an iterative perturbation approach. This optimization method systematically perturbs atomic positions within defective crystal structures, enabling a gradient-free exploration of low-energy configurations.

(I) Simulated Annealing: We employed simulated annealing, as described by Cheng and Gong¹¹², initiating optimization from defect-containing crystal structures with energies computed using ALIGNN. Random perturbations were systematically applied to atomic positions near defects, and the resulting structures' energies were re-evaluated. Lower energy structures were accepted directly, while higher energy configurations could be probabilistically accepted, controlled by a temperature parameter to avoid local minima. The temperature was progressively reduced following a predetermined schedule, ultimately guiding the optimization toward low-energy defect structures.

(II) Bayesian Optimization: Bayesian optimization¹¹² was used to efficiently identify low-energy configurations. Initially, several defect structures with small atomic perturbations were sampled, and their energies evaluated with ALIGNN. These initial samples informed a Gaussian process probabilistic model, which predicted energies and associated uncertainties for new configurations. An acquisition function balancing exploration and exploitation selected subsequent atomic configurations for evaluation. Iteratively updating the model, this method quickly converged to the lowest energy defect structures.

As an illustrative example, both simulated annealing and Bayesian optimization were applied to optimize a V_{Cd} defect within a $3 \times 3 \times 3$ CdTe supercell (**see Figure S5**). Each method rapidly identified stable defect structures within minutes, significantly faster than traditional DFT methods. Simulated annealing notably discovered a Te–Te dimer configuration, which standard DFT often overlooks without prior chemical intuition.