Supplemental material to "DeFecT-FF: A Machine Learning Force Field Framework for High Throughput Defect Modeling in CdTe-Based Solar Cells"

Md Habibur Rahman,[1)] Maitreyo Biswas,[1)] and Arun Mannodi-Kanakkithodi [1, a)]

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

Table SI Enumeration of defects across binary and alloyed Cd/Zn–Te/Se compositions. Binary compounds (CdTe, CdSe, ZnTe) contain two atomic species and therefore host 6 native defect types. Alloyed compounds contain three distinct ionic species, yielding 12 native defect types. Extrinsic defect species (Cu; As, P, N, Sb, Bi; Cl, O) contribute three defect types each (interstitial, cation substitution, anion substitution). For mixed compositions, we enforce a minimum of 10 non-equivalent positions per defect type, even through thousands of distinct symmetry-inequivalent positions exist (e.g., As_i in Cd–Se–Te). Each defect configuration is additionally perturbed using ShakeNBreak[45] to generate at least 15 symmetry-broken initial structures.

Compound	Native	Extrinsic	Native + Extrinsic	Double complexes	Total unique defects	Non-eq sites ($\times$10)	Perturbed structures ($\times$15)
CdTe	6	24	30	435	465	-	6,975
CdSe	6	24	30	435	465	-	6,975
ZnTe	6	24	30	435	465	-	6,975
$CdSe_{0.25}Te_{0.75}$	12	32	44	946	990	9,900	148,500
$CdSe_{0.50}Te_{0.50}$	12	32	44	946	990	9,900	148,500
$CdSe_{0.75}Te_{0.25}$	12	32	44	946	990	9,900	148,500
$Cd_{0.25}Zn_{0.75}Te$	12	32	44	946	990	9,900	148,500
$Cd_{0.50}Zn_{0.50}Te$	12	32	44	946	990	9,900	148,500
$Cd_{0.75}Zn_{0.25}Te$	12	32	44	946	990	9,900	148,500

Composition	PBE (Å)	HSE06 (Å)
$CdSe_xTe_{1-x}$		
CdTe ($x = 0.00$)	19.88	19.72
$CdSe_{0.25}Te_{0.75}$	19.57	19.34
$CdSe_{0.50}Te_{0.50}$	19.25	19.02
$CdSe_{0.75}Te_{0.25}$	18.94	18.71
CdSe ($x = 1.00$)	18.63	18.45
$Cd_xZn_{1-x}Te$		
CdTe ($x = 1.00$)	19.88	19.72
$Cd_{0.75}Zn_{0.25}Te$	19.56	19.35
$Cd_{0.50}Zn_{0.50}Te$	19.21	19.01
$Cd_{0.25}Zn_{0.75}Te$	18.85	18.72
ZnTe ($x = 0.00$)	18.52	18.42

Table SII Optimized supercell lattice parameters (in Å) for $CdSe_xTe_{1-x}$ and $Cd_xZn_{1-x}Te$ compounds computed from PBE and HSE06.

[a]amannodi@purdue.edu

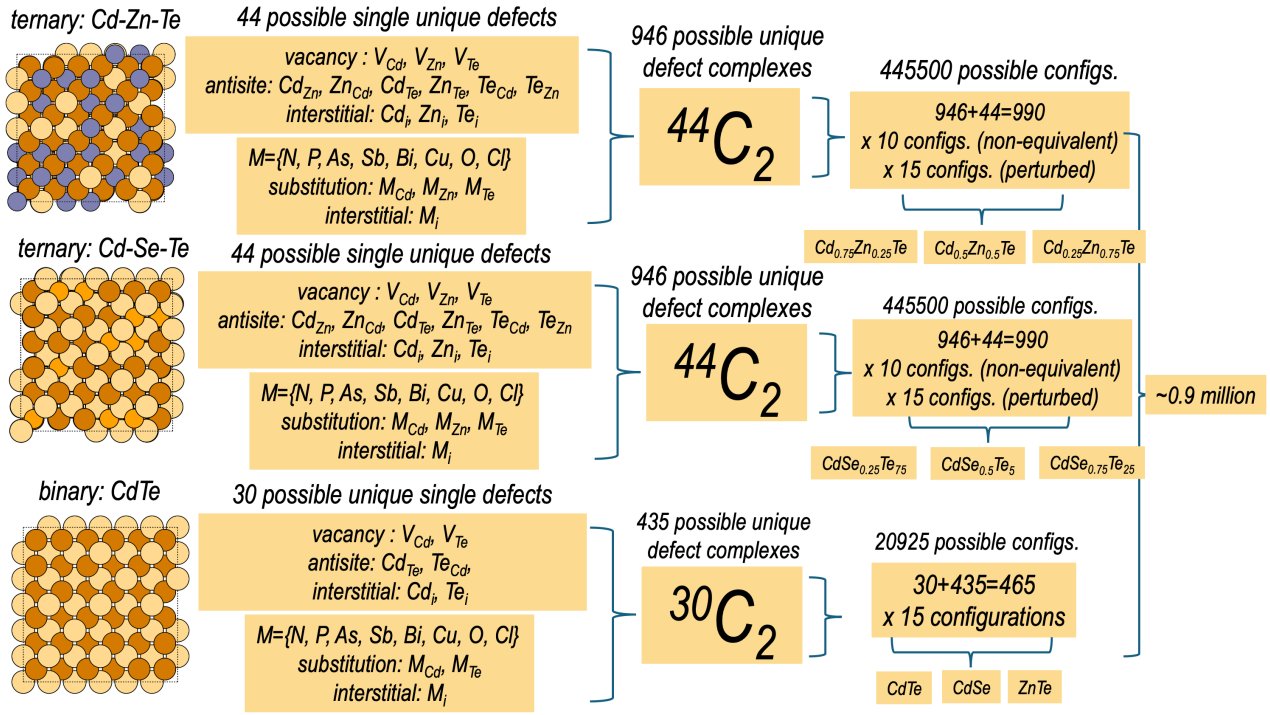

Figure S1 Schematic illustration of the combinatorial expansion of the full Cd/Zn–Te/Se defect chemical space across binary and ternary compositions. Each ternary alloy (Cd–Zn–Te and Cd–Se–Te) hosts 44 unique single-defect configurations arising from vacancies, antisites, interstitials, and eight extrinsic defects (N, P, As, Sb, Bi, Cu, O, Cl). These yield $^{44}C_2 = 946$ unique defect complexes. Combining all single defects and complexes gives 990 distinct defect types per composition, which are further expanded by sampling at least 10 non-equivalent lattice positions per defect and generating 15 ShakeNBreak[45] symmetry-broken perturbations per configuration, resulting in approximately $\sim 1.495 \times 10^5$ total structures per ternary alloy. For the binary compound CdTe (and similarly CdSe and ZnTe), 30 unique single defects generate $^{30}C_2 = 435$ complexes, yielding unique 465 defects, each further expanded through 15 perturbed initial structures to give 6,975 configurations.

Detailed Description of the DFT Datasets

The full DFT dataset used in this work includes two level of theories: a broad GGA–PBE dataset containing tens of thousands of bulk and defect configurations, and subset of PBE data refinement using the HSE06 hybrid functional. The majority of the PBE dataset was collected from previously published works from our group covering defects in a wide range of II–VI semiconductors[40,54,61,74]. These structures were combined with a significant number of new PBE calculations, particularly for alloy compositions and defect configurations that were not part of prior studies. Our chemical space spans binary Cd/Zn–Te/Se/S compounds, selected ternaries, and several quaternary alloys. **Table SIII** lists the full set of bulk compounds included in this study. Each composition is simulated using both $2 \times 2 \times 2$ (64-atom) zincblende supercells generated using the special quasirandom structure (SQS) approach[87]. For each composition, the mixing fraction x is systematically varied in increments of 0.125 (i.e., $n/8$ for $n \in \{0, \ldots, 8\}$), yielding a total of 81 unique compositions[61].

Table SIII Summary of bulk compounds included in the PBE/HSE dataset. The full set spans binary, ternary, and quaternary compositions in the Cd/Zn–Te/Se/S chemical space.

Category	Compounds
Binary	CdTe, CdSe, CdS, ZnTe, ZnSe, ZnS
Ternary alloys	$CdSe_xTe_{1-x}$, CdS_xSe_{1-x}, $ZnSe_xTe_{1-x}$, ZnS_xSe_{1-x}
Quaternary alloys	$Cd_xZn_{1-x}S$, $Cd_xZn_{1-x}Se$, $Cd_xZn_{1-x}Te$, $Cd_{0.5}Zn_{0.5}S_xSe_{1-x}$, $Cd_{0.5}Zn_{0.5}Se_xTe_{1-x}$

Across all compositions and supercell sizes, the PBE dataset contains more than 10,000 bulk structures, including not only the fully relaxed configurations but also all intermediate snapshots along the geometry-optimization trajectories.

Dataset	Supercell Size	Data Points
Bulk dataset from Cd/Zn-Te/Se/S binary, ternary and quaternary alloys	2×2×2	10080
Bulk dataset from $CdSe_xTe_{1-x}$ alloys	3×3×3	26
Defect dataset from 6 Cd/Zn-Te/Se/S binary compounds	2×2×2	7302 (q=+2), 6201 (q=+1), 8203 (q=0), 6361 (q=-1), 7689 (q=-2)
Defect dataset from $CdSe_xTe_{1-x}$ and $Cd_xZn_{1-x}Te$ alloys	3×3×3	594 (q=+2), 533 (q=+1), 643 (q=0), 514 (q=-1), 560 (q=-2)
Defect dataset from CdTe/ZnTe interface	3×3×6	375 (q=+2), 380 (q=+1), 401 (q=0), 381 (q=-1), 330 (q=-2)
Defect dataset from CdTe dislocation core	–	210 (q=+2), 223 (q=+1), 263 (q=0), 220 (q=-1), 231 (q=-2)

Table SIV 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.

Dataset	Supercell Size	Data Points
Bulk dataset from Cd/Zn-Te/Se/S binary compounds and alloys	2×2×2	5400
Defect dataset from 6 binary compounds	2×2×2	4302 (q=+2), 4201 (q=+1), 6203 (q=0), 4361 (q=-1), 4689 (q=-2)
Defect dataset from $CdSe_xTe_{1-x}$ and $Cd_xZn_{1-x}Te$ alloys	3×3×3	371 (q=+2), 333 (q=+1), 402 (q=0), 321 (q=-1), 350 (q=-2)

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

Each structure is labeled by its crystal formation energy (CFE), defined for a general composition $Cd_aSe_bTe_c$ as:

$$CFE = \frac{E(Cd_aSe_bTe_c) - aE(Cd) - bE(Se) - cE(Te)}{N_{atoms}}. \quad (3)$$

Here, $E(Cd_aSe_bTe_c)$ 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 $CdSe_xTe_{1-x}$ composition. $E(Cd)$, $E(Se)$, and $E(Te)$ are respectively the per-atom energies of Cd, Se, and Te in their known elemental standard states, and $N_{atoms} = a+b+c$ is the total number of atoms in the supercell.

The defect dataset contains native vacancies, interstitials, antisites, extrinsic interstitials and substitutional dopants (Cu, As, P, N, Sb, Cl, O), and selected complexes across multiple compositions. All defect calculations were performed in either $2 \times 2 \times 2$ or $3 \times 3 \times 3$ supercells, and multiple charge states were considered. Additionally, we incorporated CdTe dislocation–core structural models derived from prior STEM-based work[81]. These structures correspond to a 4.8° low-angle (110)‖(110) tilt boundary, in which the extended defect is represented by a periodic array of Lomer-type dislocation cores. Two experimentally observed core reconstructions were simulated: a *Type-I* core, where the absence of a central Cd column produces a Te-terminated dislocation core with broken Te–Cd bonds that generate deep mid-gap states; and a *Type-II* core, where a restored central Cd column leads to a distinct mixed Cd–Te coordination environment and an electronic state. Together, these two structures capture the essential bonding rearrangements, dangling-bond motifs, and charge perturbations intrinsic to realistic CdTe dislocation cores identified through atomic-resolution STEM imaging[81]. We also constructed a CdTe–ZnTe heterointerface by merging fully relaxed 3×3×3 supercells of CdTe and ZnTe. Small in-plane strains were applied to lattice-match the two materials, followed by volume-conserving structural relaxation to obtain a physically consistent interface. For all new defect calculations in the 3×3×3 binary, alloyed, and interface the `Doped` package[88] was used to generate symmetry-broken initial configurations via bond distortions and atomic rattling. In the unified dataset, the label `bulk` denotes pristine supercells; `defect` refers to supercells containing vacancies, interstitials, antisites, dopants, or multi-defect complexes; `interface` corresponds to CdTe–ZnTe heterostructures containing defects at or near the interfacial region; and `dislocation_core` designates structures based on the Type-I and Type-II STEM-derived dislocation-core models, with defects introduced within their reconstructed core environments. The complete PBE dataset distributions used for ALIGNN training are reported in **Figure S2**, **Figure S3**, and **Table SIV**.

A curated subset of bulk and defect structures from the PBE dataset was re-optimized using the HSE06 hybrid functional with a mixing parameter of $\alpha = 0.25$. Due to high computational costs, HSE06 calculations were performed only for representative configurations spanning different compositions, and chemical environments. For defect calculations, lattice parameters were updated to the HSE06-optimized bulk volume before relaxation. All hybrid-functional calculations used Γ-point sampling and reduced plane-wave cutoffs appropriate for large supercells. The distributions of the complete HSE dataset are presented in **Figure S4** and **Table SV**. This dataset forms the basis for the MLFF[76] models used in this work.

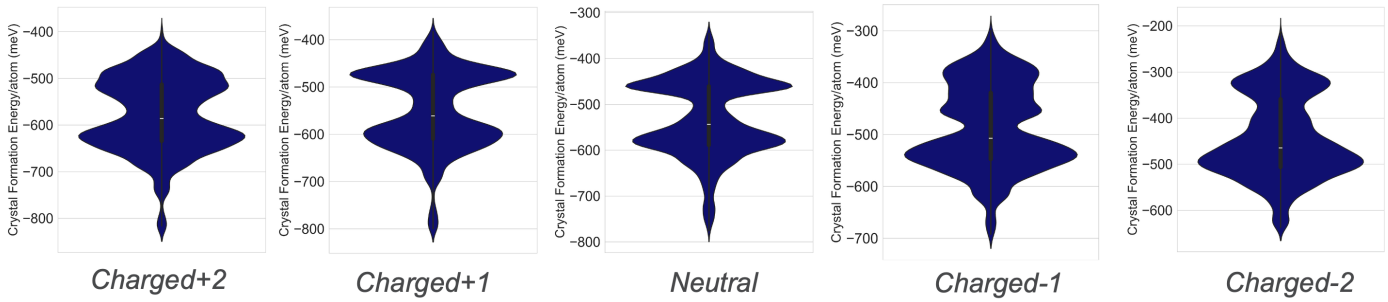

Figure S2 Violin distributions of crystal formation energy per atom (meV) for charge states $+2$, $+1$, 0 (neutral), -1, and -2 in the GGA-PBE dataset.

Graph neural network models for direct prediction of crystal formation energy

ALIGNN was developed by Choudhary *et al.*[76] 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 (PBE) 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 S5**, in terms of ALIGNN-predicted CFE vs DFT-computed CFE for only the test set data points. Models pictured in **Figure S5(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 S5(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 S5(c)**. For the defect-only ALIGNN model in **Figure S5(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 intersitial 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[52,54,77] 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[48,49,76,109–113], 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.

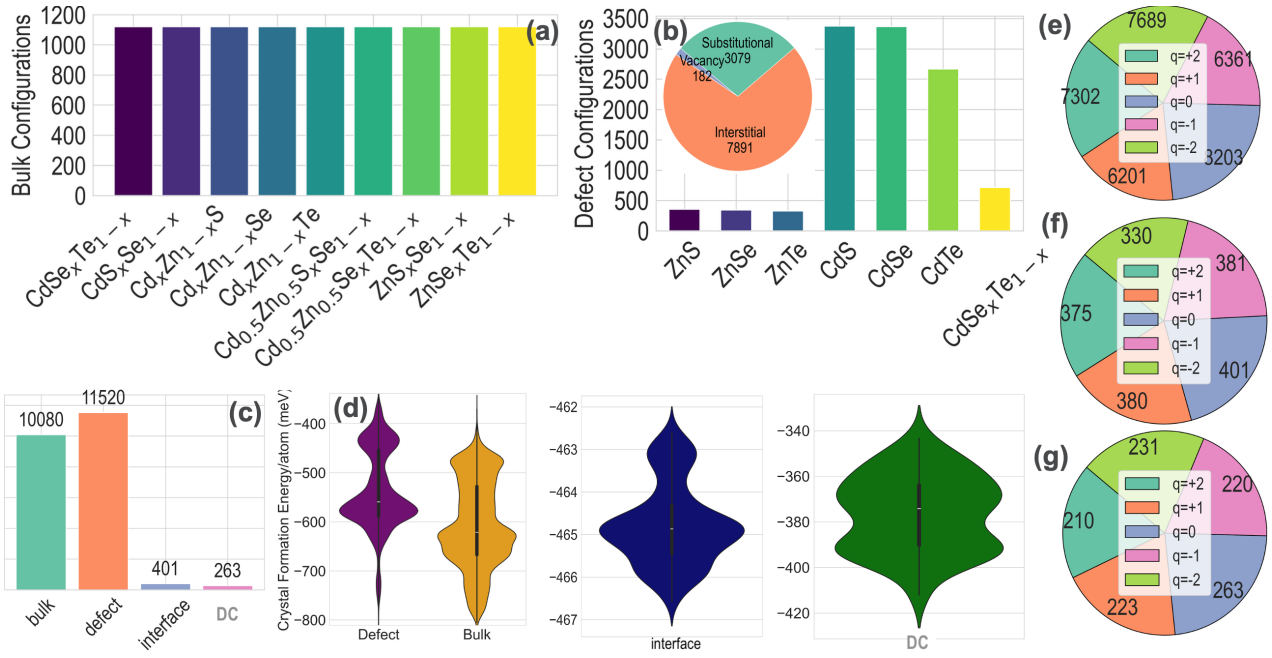

Figure S3 Statistics of the GGA-PBE dataset: (a) Number of bulk configurations corresponding to $CdSe_xTe_{1-x}$, CdS_xSe_{1-x}, $Cd_xZn_{1-x}S$, $Cd_xZn_{1-x}Se$, $Cd_xZn_{1-x}Te$, $Cd_{0.5}Zn_{0.5}S_xSe_{1-x}$, $Cd_{0.5}Zn_{0.5}Se_xTe_{1-x}$, ZnS_xSe_{1-x}, and $ZnSe_xTe_{1-x}$ compositions. (b) Number of neutral defect configurations in CdS, CdSe, CdTe, ZnS, ZnSe, ZnTe, and different $CdSe_xTe_{1-x}$ 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 dislocation core (DC) configurations in the dataset. (d) Violin plots showing the distribution of crystal formation energy across all the defect, bulk, interface and DC 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 DC configurations, respectively.

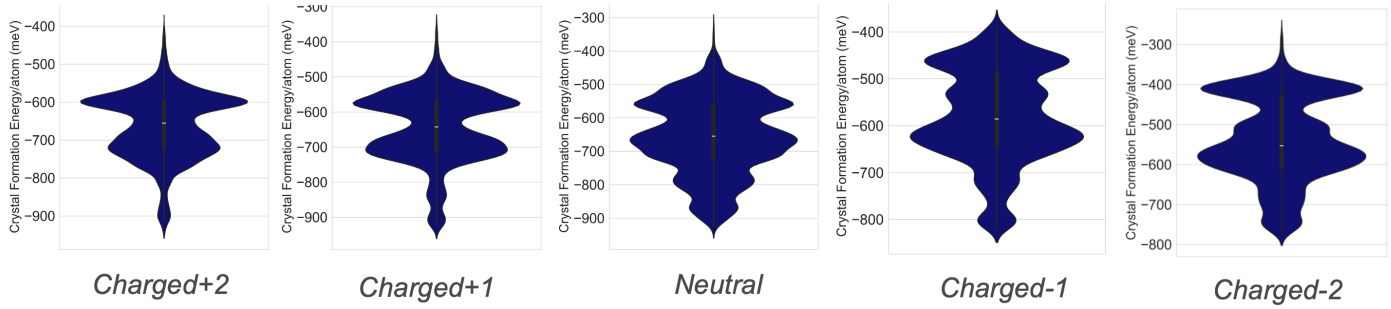

Figure S4 Violin distributions of crystal formation energy per atom (meV) for charge states $+2$, $+1$, 0 (neutral), -1, and -2 in the HSE06 dataset.

3. An ALIGNN model trained exclusively on defects in 2×2×2 supercell structures and then used to predict the CFE of 3×3×3 supercell defect structures.

ALIGNN shows poor transferability when trained only on bulk data and used to predict for defects; as shown in **Figure S6(a)**, the prediction RMSEs range from $\sim$ 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 S6(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 S3(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 S5** and there are some very clear outliers, as pictured in **Figure S6(b)**. Lastly, when the model is trained on only 2×2×2 supercell structures and used to predict for 3×3×3 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.

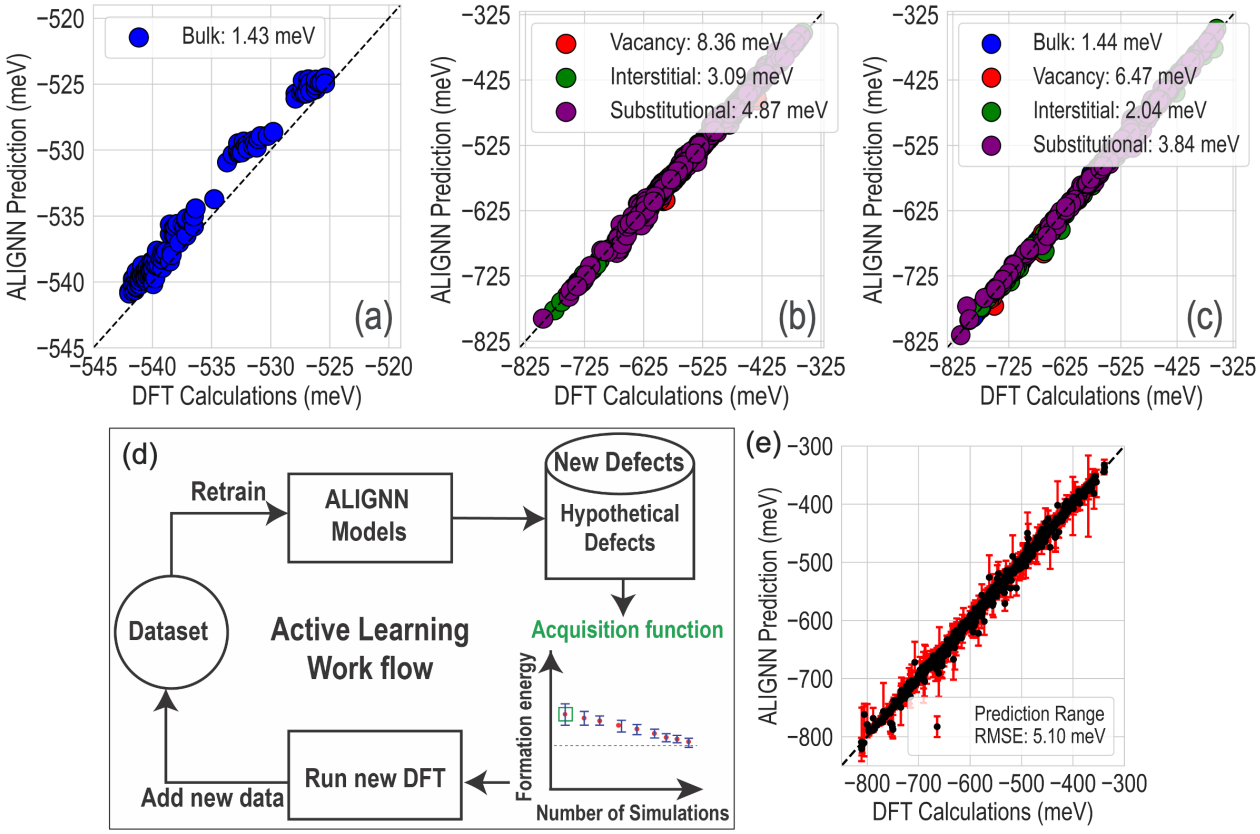

Figure S5 Parity plots for ALIGNN models trained on the GGA dataset using (a) only bulk structures, (b) only defect structures, with predictions distinguished in terms of type of defect, and (c) both bulk and defect structures. (d) Active learning (AL) workflow implemented in this work: standard deviation 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.

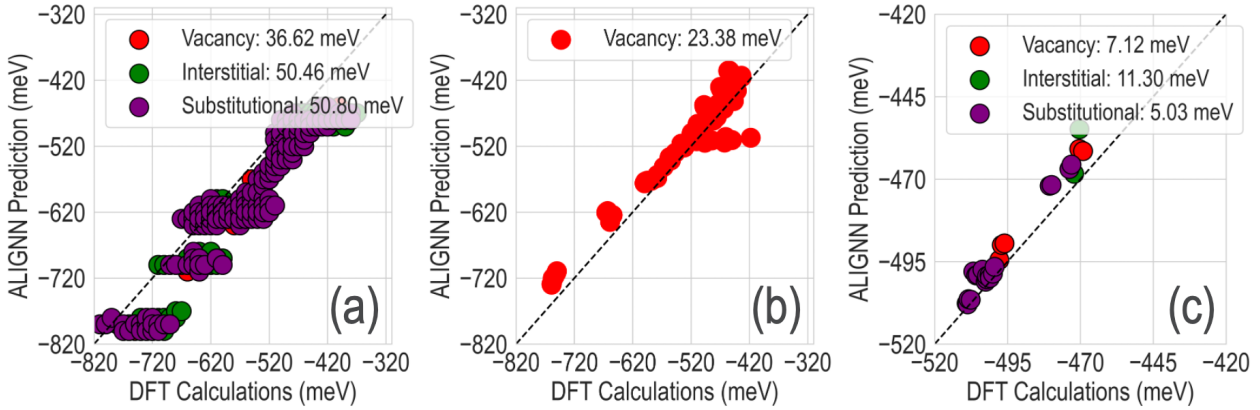

Figure S6 Parity plots comparing ALIGNN predictions to DFT calculations under different training conditions: (a) ALIGNN (trained solely on bulk structures) predictions vs. DFT calculations for the defect dataset. (b) ALIGNN (trained exclusively on interstitial defect structures) predictions vs. DFT calculations for the vacancy dataset. (c) ALIGNN (trained only on 2×2×2 supercell bulk and defect structures) predictions vs. DFT calculations for defects in a 3×3×3 supercell.

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.

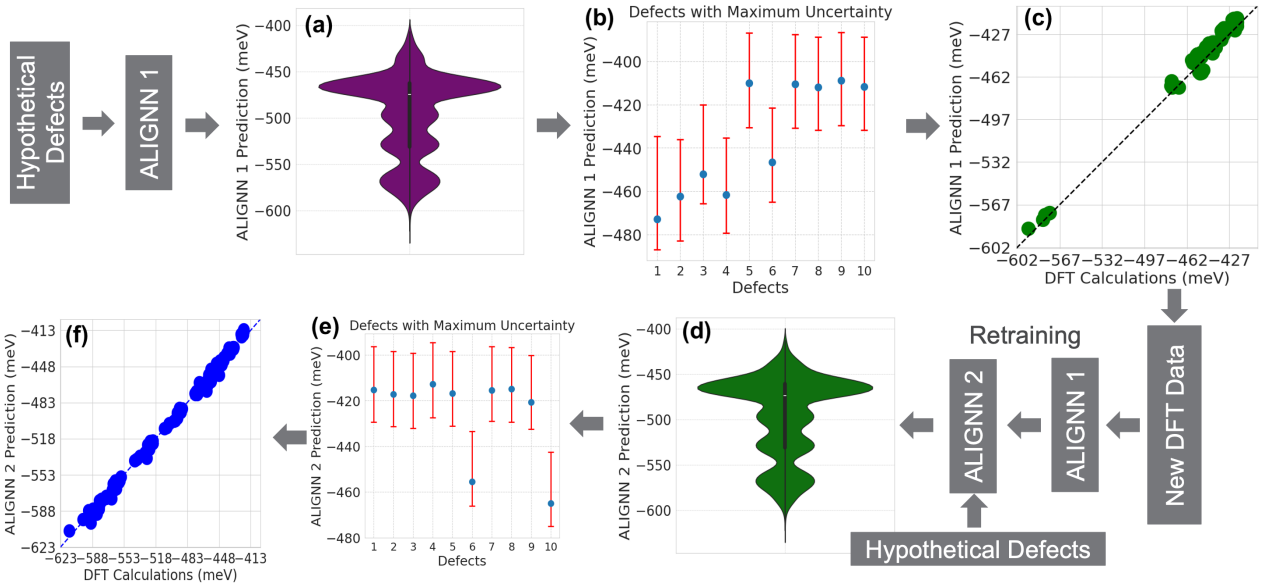

Figure S7 (a) Violin plot showing the mean crystal formation energy (CFE) predicted by the initial ALIGNN-1 models, averaged across 100 models. (b) Defects with maximum uncertainties identified through ALIGNN-1 models. (c) Comparison of ALIGNN-1 predictions vs DFT calculations for 200 selected defects. (d) Violin plot of mean CFE from ALIGNN-2 models. (e) Defects with maximum uncertainties identified through ALIGNN-2 models. (f) Comparison of ALIGNN 2 predictions vs DFT calculations for the 200 selected defects.

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 S5(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 S7(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 function[71]. The MU criterion is defined as $\mathrm{MU}(x) = \sigma(x)$, where $\sigma(x)$ denotes the standard deviation (uncertainty) of the prediction. **Figure S7(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 S7(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 S7(d)**. The predictions are again evaluated using the acquisition function as shown in **Figure S7(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 S7(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, 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.

Detailed Description of ALIGNN-Based Geometry Optimization

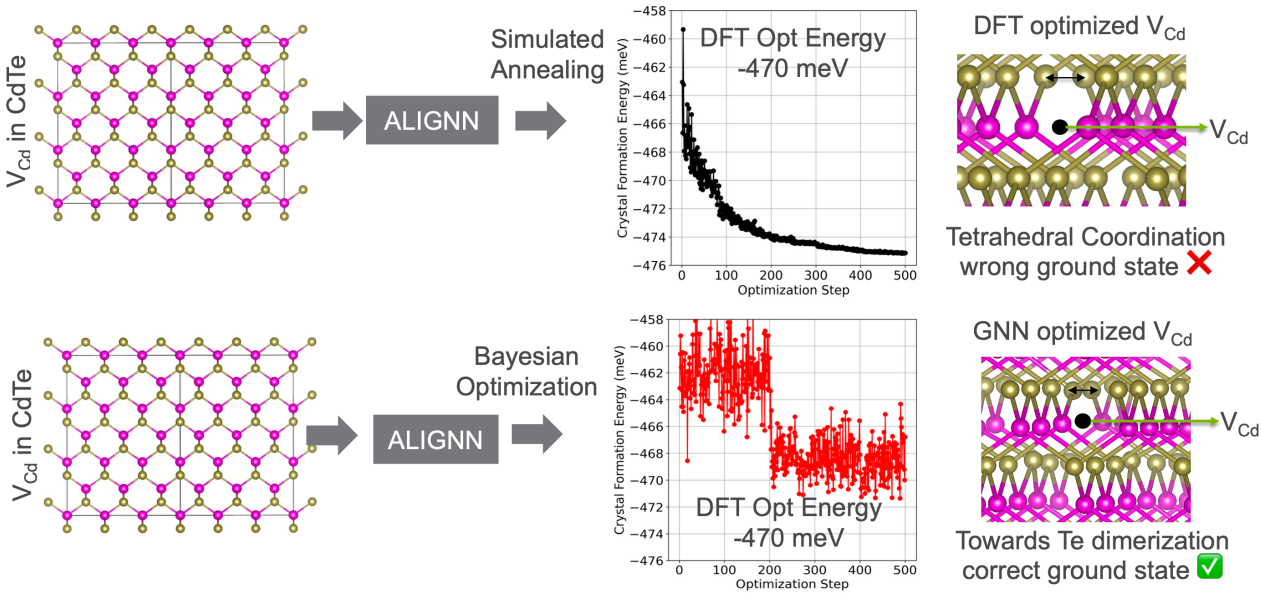

Figure S8 Optimization of Cd vacancy (V_{Cd}) in CdTe using the trained ALIGNN model with two different optimization strategies: simulated annealing and Bayesian optimization.

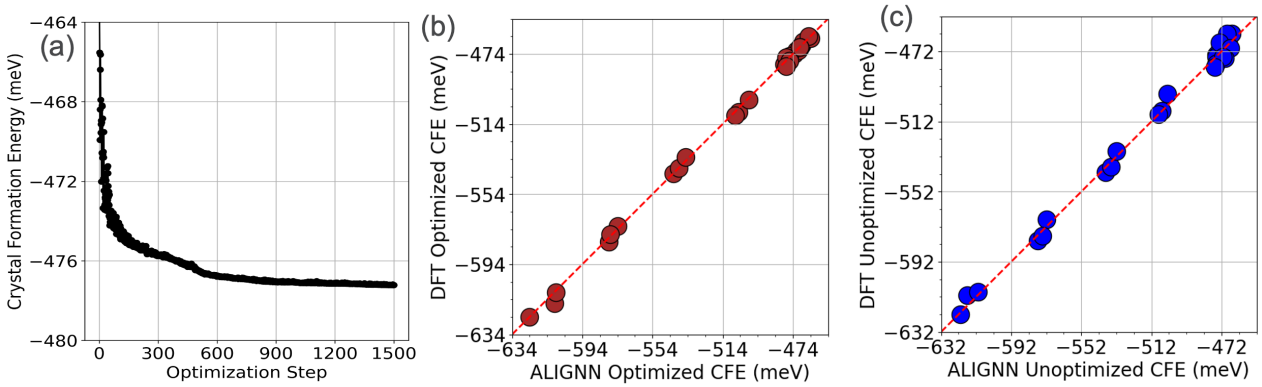

Figure S9 a) Simulated annealing optimization of a Cl_{Te} defect in $CdSe_{0.50}Te_{0.50}$ using ALIGNN-predicted crystal formation energy (CFE), showing energy minimization over successive steps. (b) Parity plot comparing ALIGNN-optimized CFE with DFT-optimized CFE across a set of defect structures, demonstrating strong agreement. (c) Parity plot comparing ALIGNN-unoptimized CFE with DFT-unoptimized CFE.

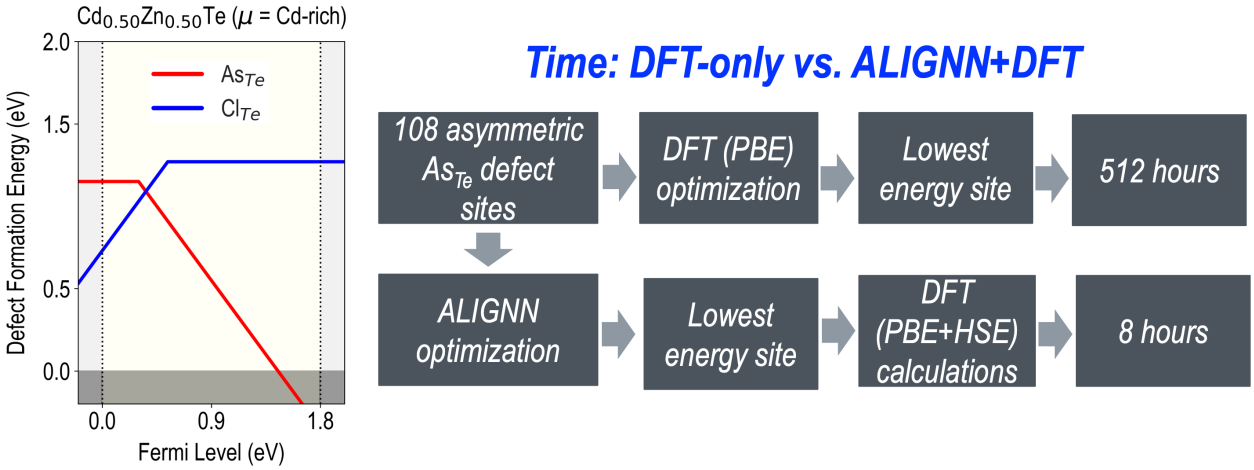

Figure S10 Comparison of computational efficiency between a DFT-only workflow and the ALIGNN+SA+DFT approach for evaluating Cl_{Te} and As_{Te} defects in $Cd_{0.50}Zn_{0.50}Te$ under Cd-rich conditions. Left: Defect formation energies (DFEs) as a function of Fermi level, computed using the HSE06+SOC 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.

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

1. Simulated Annealing (SA)[114], which is a gradient-free stochastic method employing random atomic perturbations guided by an annealing schedule.

2. Bayesian Optimization (BO)[114], which uses Gaussian processes to iteratively identify low energy configurations through atomic displacement exploration.

As an illustrative example, both simulated annealing and Bayesian optimization were applied to optimize a V_{Cd} defect within a 3×3×3 CdTe supercell **(see Figure S8)**. 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.

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 3×3×3 CdTe supercell, and the results are pictured in **Figure S8**. 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.[20] 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 $CdSe_xTe_{1-x}$ and $Cd_xZn_{1-x}Te$ chemical space as shown in **Figure S9**. For example, Cl_{Te} in $Cd_{0.50}Zn_{0.50}Te$ was successfully optimized using ALIGNN+SA, as shown in **Figure S9 (a)**.

Table SVI Performance of pretrained machine-learning force-field (MLFF) models applied directly to the PBE test dataset without retraining. The reported errors correspond to the root-mean-square error (RMSE) in crystal formation energy (CFE).

Model	Pretrained Version	RMSE (meV/atom)
MACE[79]	`MACE_MPtrj_2022.9.model`	93.79
M3GNet[53]	`M3GNet-MatPES-PBE-v2025.1-PES`	62.10
CHGNet[80]	`CHGNet-MatPES-PBE-2025.2.10-2.7M-PES`	59.50

Figure S10 illustrates the DFEs of As_{Te} and Cl_{Te} in $Cd_{0.50}Zn_{0.50}Te$ 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-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.

PBE Machine learning Force Field (MLFF) Model

M3GNet-based[53] MLFF models were trained using the energies, atomic forces, and stresses extracted from GGA–PBE calculation trajectories. 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[115,116] 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 S11 (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 S12**. 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 dislocation core (DC). 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. Dislocation core defect structures show slightly larger errors of $>$ 9.1 meV/atom, which is expected given their more complex local environments and atomic rearrangements. 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 S13**. For example, we optimized the $As_{Te}+Cl_{Te}$ defect in $CdSe_{0.12}Te_{0.88}$ 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 S11(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} substitutional defect in a CdTe dislocation core 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.

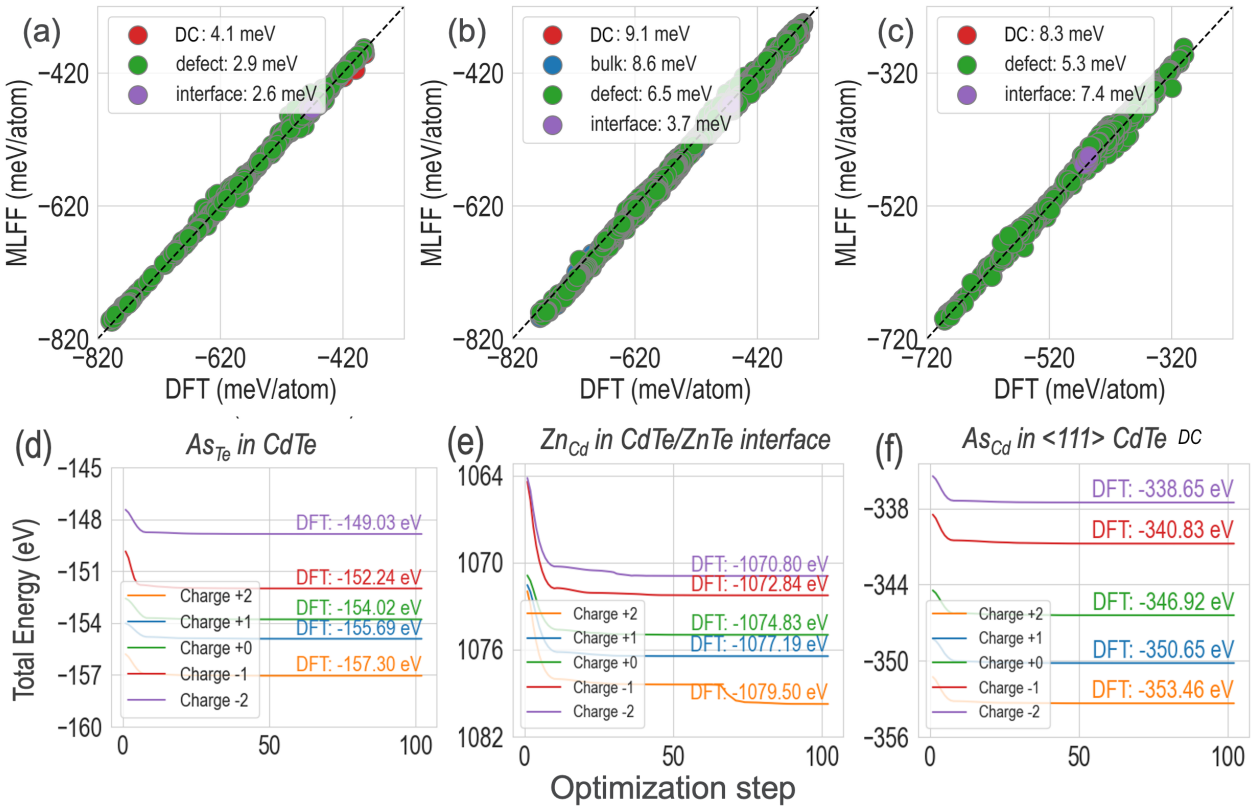

Figure S11 Performance of M3GNet-MLFF models trained on the GGA-PBE dataset, shown in terms of predicted vs DFT crystal formation energy parity plots. The models were trained separately for (a) defect configurations with charge *q=+1*, (b) neutral *q=0* defect and bulk configurations, and (c) defect 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 "DC" indicates defects situated in CdTe dislocation core 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 dislocation core structure (DC).

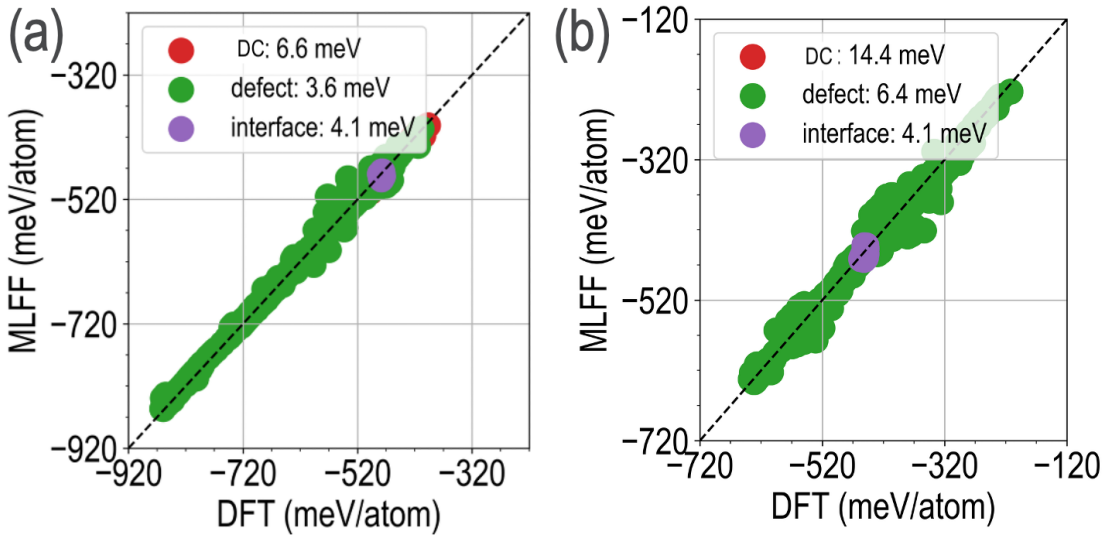

Figure S12 Parity plots for M3GNet-MLFF models trained on the GGA dataset, shown in terms of predicted vs actual (from DFT) crystal formation energies, trained separately for (a) defect configurations with charge *q=+2*, (b) defect configurations with charge *q=-2*. Here, "defect" represents bulk supercells containing a point defect or defect complex, "interface" corresponds to defects located at CdTe-ZnTe interfaces, and "DC" indicates defects situated at CdTe dislocation core.

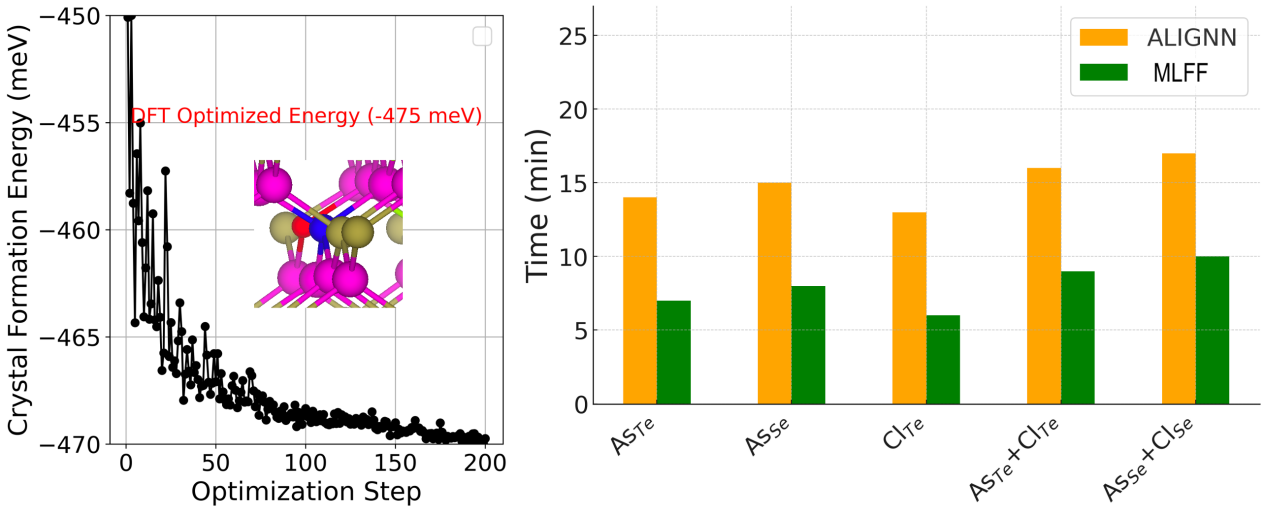

Figure S13 (a) Optimization of an As_{Te} + Cl_{Te} complex in $CdSe_{0.12}Te_{0.88}$ using an M3GNET-MLFF model, and (b) comparison of the time taken by ALIGNN and the MLFF for optimizing selected defects in $CdSe_{0.12}Te_{0.88}$.

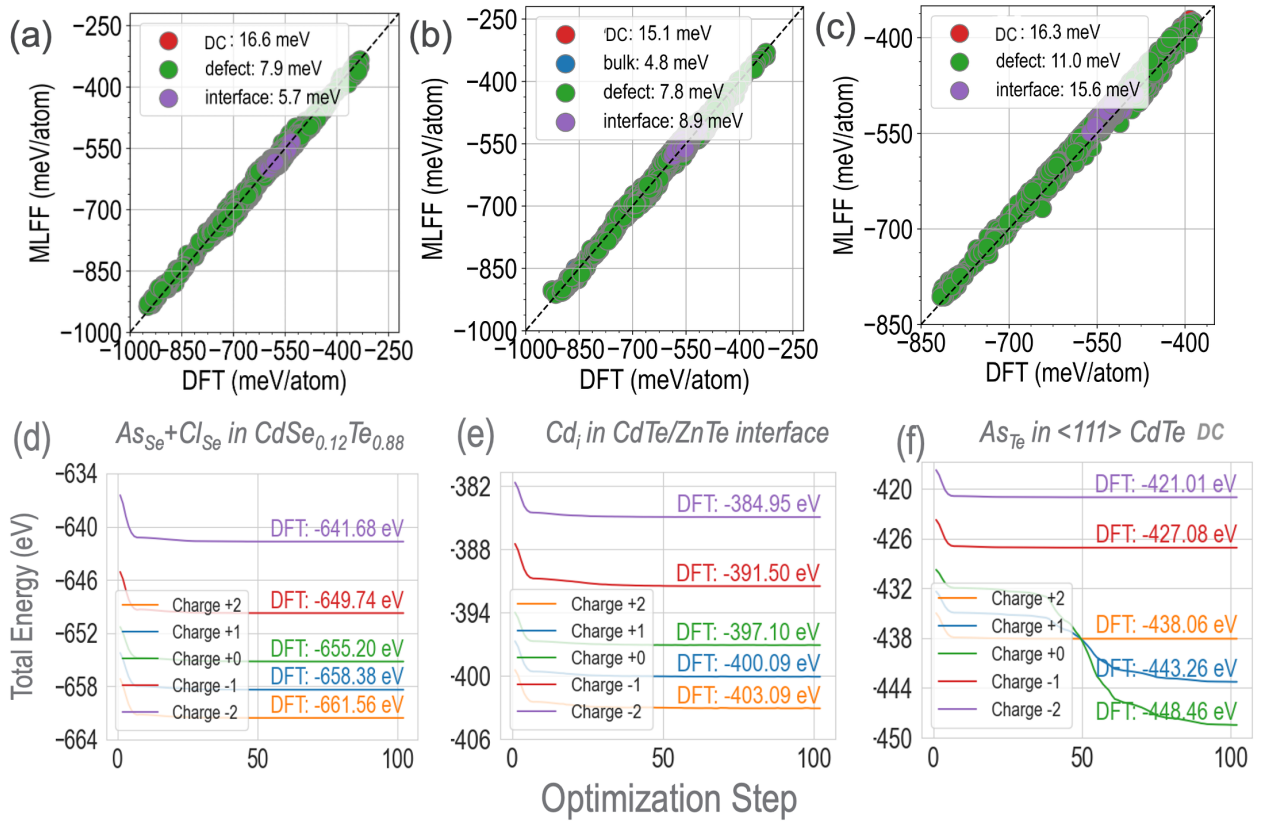

Figure S14 Performance of the M3GNet-MLFF models trained on the augmented HSE06 dataset, shown in terms of predicted vs DFT crystal formation energy parity plots. The models were trained separately for (a) defect configurations with charge *q=+1*, (b) neutral *q=0* defect and bulk configurations, and (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 "DC" indicates defects situated at CdTe dislocation core. Plots in (d–f) show the geometry optimization process taking into account 5 charge states for different example defect configurations: (d) an As_{Se}+Cl_{Se} defect complex in $CdSe_{0.12}Te_{0.88}$, (e) a Cd_i defect at the CdTe/ZnTe interface, and (f) As_{Te} in <111> CdTe dislocation core structure.

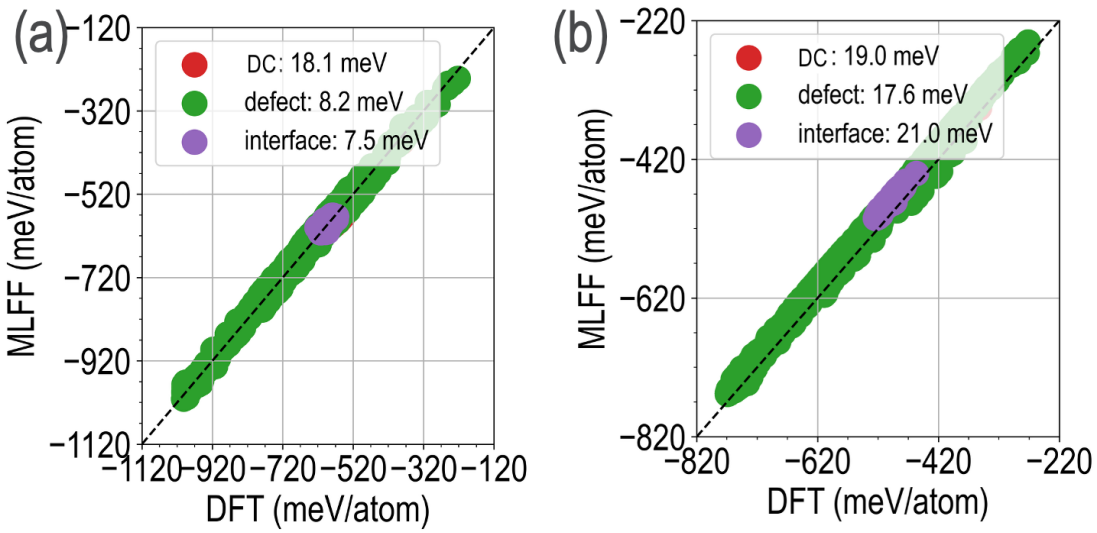

Figure S15 Parity plots for M3GNet-MLFF models trained on the HSE06 dataset, shown in terms of predicted vs actual (from DFT) crystal formation energies, trained separately for (a) defect configurations with charge *q=+2*, (b) defect configurations with charge *q=-2*. Here, "defect" represents bulk supercells containing a point defect or defect complex, "interface" corresponds to defects located at CdTe-ZnTe interfaces, and "DC" indicates defects situated at CdTe dislocation core structure.

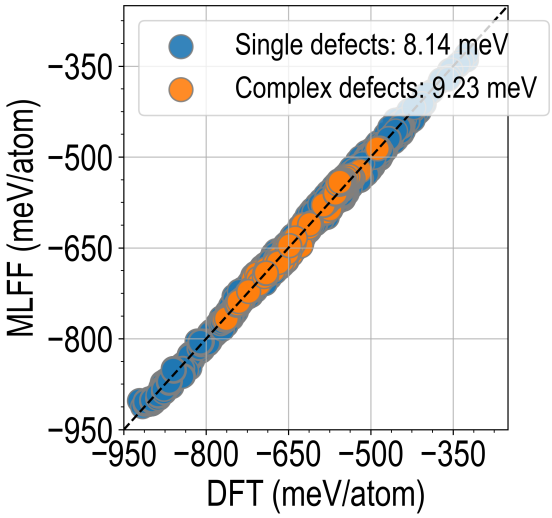

Figure S16 Parity plot comparing DFT and MLFF-predicted crystal formation energies (CFE) for single and complex defects (q=0 charge state).

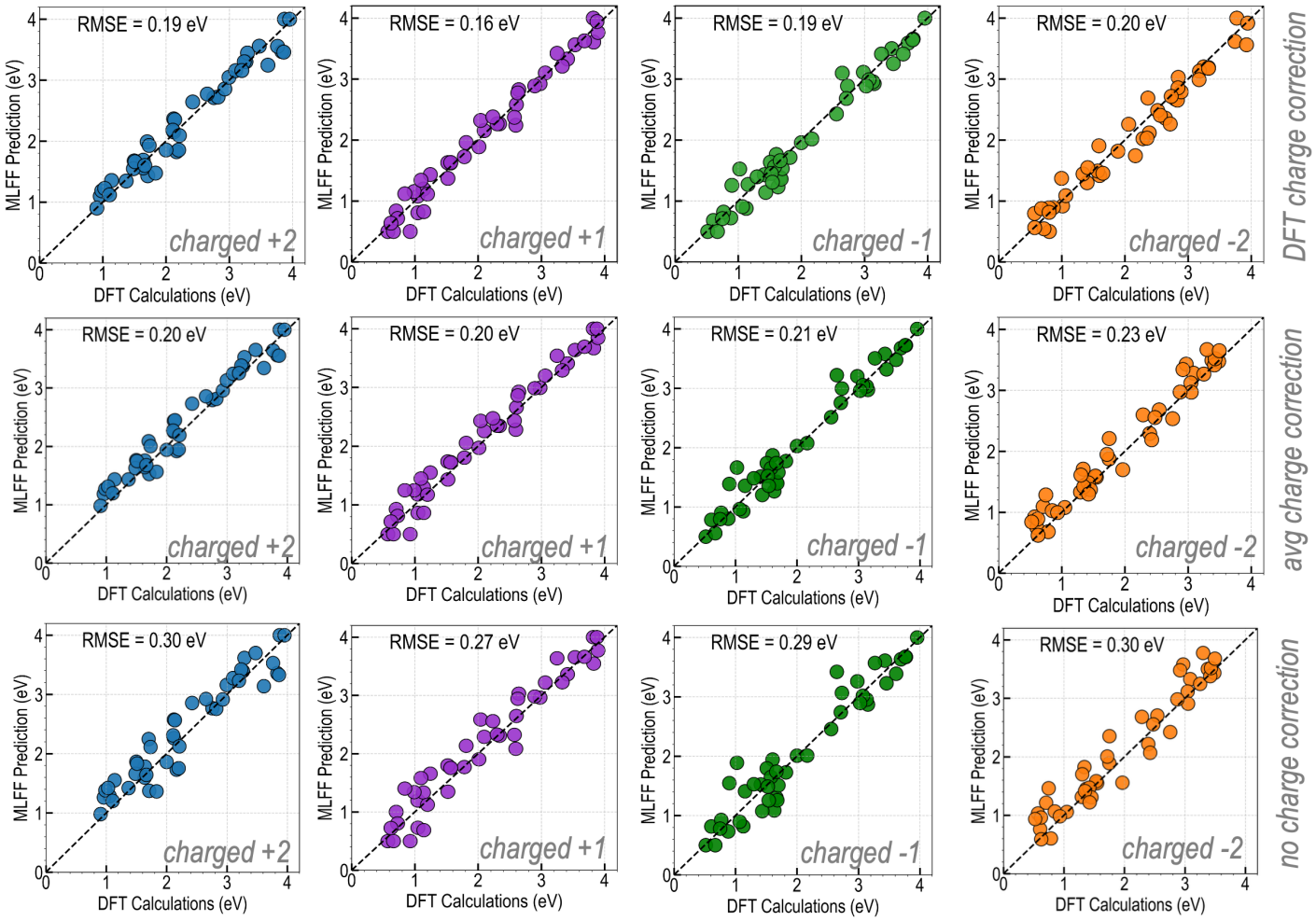

Figure S17 Parity plots comparing MLFF-predicted defect formation energies with DFT values for different charge states and charge–correction schemes. Each panel shows the MLFF prediction plotted against the corresponding DFT value for charged defects with $q = +2$, $+1$, -1, and -2. The first row uses actual DFT charge corrections, the second row uses averaged charge corrections for each charge, and the third row applies no charge correction.

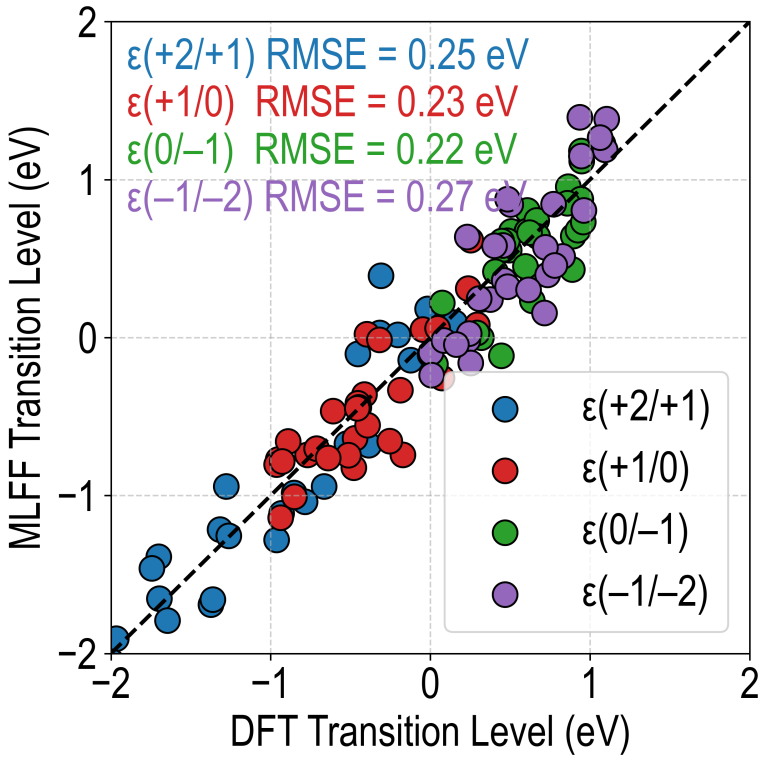

Figure S18 Parity plot comparing MLFF-predicted and DFT-calculated charge-transition levels, with average charge correction value applied to MLFF prediction in each charge state. Four types of transition levels are calculated, namely +2/+1, +1/0, 0/-1, and -1/-2.

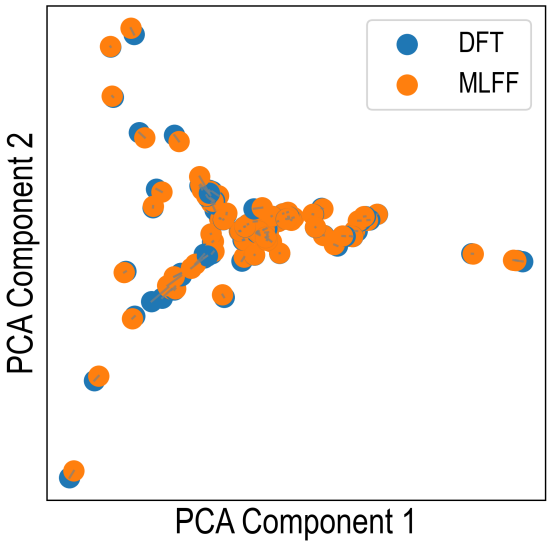

Figure S19 Comparison of DFT- and MLFF-relaxed structures using PCA on SOAP descriptors[84,85]. Each point represents the structural fingerprint of a relaxed configuration projected onto a two-dimensional PCA space. Blue and orange markers correspond to DFT and MLFF optimizations, respectively. The near-overlapping distribution of DFT and MLFF points demonstrates that the latter accurately reproduces DFT-level structural relaxations across diverse bulk and defect configurations.

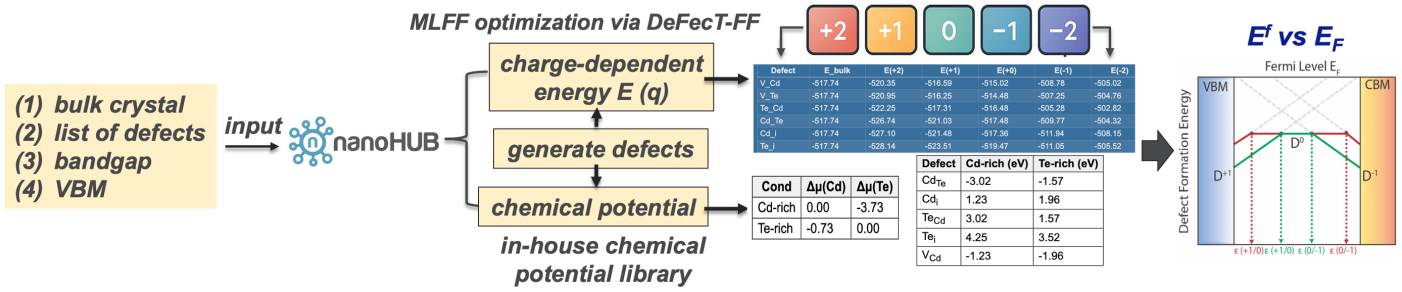

Figure S20 The `DeFecT-FF` tool takes as input the bulk crystal, list of defects, bandgap, and VBM, performs MLFF-based geometry optimization, leverages an in-house chemical potential library to calculate charge-dependent energies $E(q)$ and defect formation energies E_f as a function of the Fermi level E_F, and finally constructs E_f–E_F diagrams for defect thermodynamics. This tool is accessible via a nanoHUB web application.

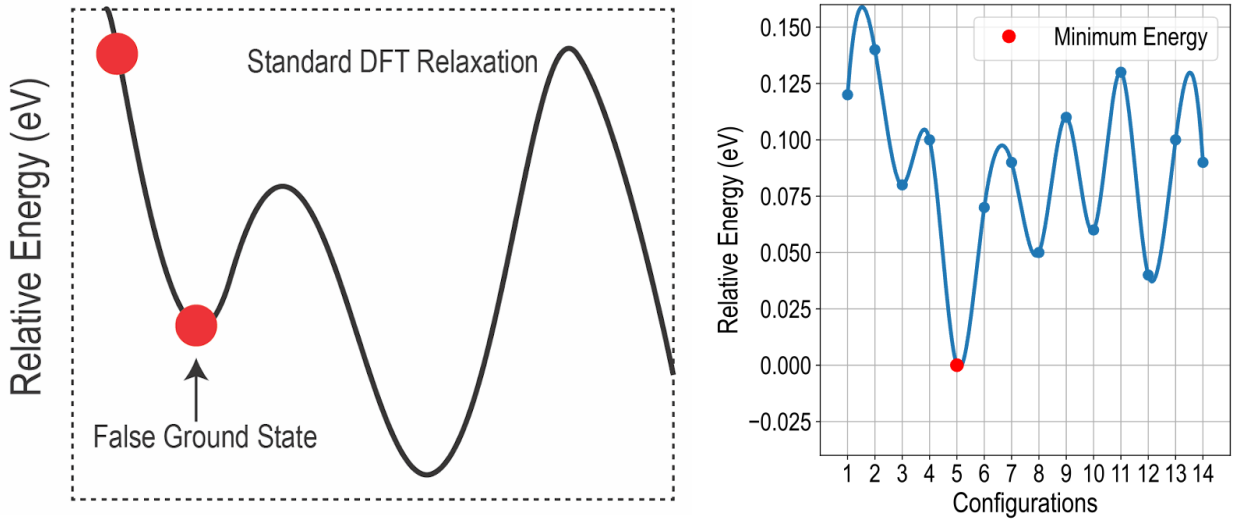

Figure S21 Illustration of the configuration search for As_{Se} defects. The left panel shows how a standard DFT relaxation can become trapped in a false ground state, while the right panel presents the relative energies of 14 symmetry-broken configurations obtained using the HSE-based MLFF. The true lowest-energy configuration (configuration 5) is highlighted in red.

Table SVII Comparison of RMSE (eV) in MLFF defect formation energies for three correction schemes: (i) no charge correction, (ii) average charge-offset correction, and (iii) DFT charge correction applied to MLFF predictions.

Charge State	No Correction	Average Offset Correction	DFT Charge Correction
$q=-2$	0.30	0.23	0.20
$q=-1$	0.29	0.21	0.19
$q=0$	0.16	0.16	0.16
$q=+1$	0.27	0.20	0.16
$q=+2$	0.30	0.20	0.19

Table SVIII RMSE for crystal formation energy prediction broken down by defect type for the neutral ($q=0$) charge state.

Defect Type	RMSE (meV/atom)
Vacancy	9.5
Antisite	8.1
Substitutional	8.6
Interstitial	7.6

Table SIX Composition-resolved test RMSE for neutral crystal formation energy (CFE) prediction across bulk and defect configurations in all compounds in the DeFecT-FF chemical space. Binary compounds show slightly lower errors compared to ternary alloys due to reduced compositional disorder.

Compound	RMSE (meV/atom)
CdTe	7.2
CdSe	7.5
ZnTe	7.8
$CdSe_{0.25}Te_{0.75}$	8.6
$CdSe_{0.50}Te_{0.50}$	9.1
$CdSe_{0.75}Te_{0.25}$	9.4
$Cd_{0.25}Zn_{0.75}Te$	8.9
$Cd_{0.50}Zn_{0.50}Te$	9.3
$Cd_{0.75}Zn_{0.25}Te$	8.4

Table SX Out-of-distribution (OOD) and in-distribution (ID) test results for crystal formation energy prediction. OOD RMSE corresponds to compositions not included in the MLFF training set, while ID RMSE corresponds to the same compositions after being added to the training set, demonstrating the improvement in predictive accuracy upon inclusion.

Composition	OOD RMSE (meV/atom)	ID RMSE (meV/atom)
$CdSe_{0.12}Te_{0.88}$	12.4	8.3
$CdSe_{0.06}Te_{0.94}$	12.8	8.7

Charge Correction Using the Freysoldt Scheme

Charged defects in periodic boundary conditions introduce spurious electrostatic interactions between the charged defect, its periodic images, and the compensating background charge. To correct for these finite-size effects, we employed the Freysoldt[17] correction scheme as implemented in the `sxdefectalign` code. The correction energy consists of two components: an image-charge term (E_{PC}) and a potential alignment term ($q\Delta V$):

$$E_{\mathrm{corr}} = E_{\mathrm{PC}} + q\Delta V. \tag{4}$$

The image-charge correction is estimated using an isotropic model charge distribution screened by the static dielectric constant ε of the host material:

$$E_{\mathrm{PC}} = \frac{q^2 \alpha}{2 \varepsilon L}, \tag{5}$$

where q is the defect charge state, α is the Madelung constant for the supercell geometry, and L is the effective lattice parameter of the supercell. The potential alignment term ΔV is obtained from the planar-averaged electrostatic potential difference between the defective and bulk supercells in the far-field region:

$$\Delta V = V_{\mathrm{defect}}^{\mathrm{far}} - V_{\mathrm{bulk}}. \tag{6}$$