npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01303-9

Machine-learning structural reconstructions for accelerated point defect calculations

Check for updates

Irea Mosquera-Lois[1], Seán R. Kavanagh[1], Alex M. Ganose[2] & Aron Walsh[1,3]

Defects dictate the properties of many functional materials. To understand the behaviour of defects and their impact on physical properties, it is necessary to identify the most stable defect geometries. However, global structure searching is computationally challenging for high-throughput defect studies or materials with complex defect landscapes, like alloys or disordered solids. Here, we tackle this limitation by harnessing a machine-learning surrogate model to qualitatively explore the structural landscape of neutral point defects. By learning defect motifs in a family of related metal chalcogenide and mixed anion crystals, the model successfully predicts favourable reconstructions for unseen defects in unseen compositions for 90% of cases, thereby reducing the number of first-principles calculations by 73%. Using $CdSe_xTe_{1-x}$ alloys as an exemplar, we train a model on the end member compositions and apply it to find the stable geometries of all inequivalent vacancies for a range of mixing concentrations, thus enabling more accurate and faster defect studies for configurationally complex systems.

Defects control the properties of many functional materials and devices[1], like solar cells[2,3], batteries[4,5], catalysts[6–8], and quantum computers[9–12]. To discover better materials for these applications it is thus necessary to predict how their defects behave. However, defect calculations are computationally demanding. The large supercells and high level of theory required to obtain robust predictions typically limit point defect analysis to in-depth studies of specific materials. In a move towards data-driven defect workflows[13], defect databases[14–20] and surrogate models have been developed to predict defect properties, like the dominant defect type[18], formation[19,21–35] and migration[35] energies, and charge transition levels[19,25,36]. By learning the relationship between defect structure and properties, these models enable high-throughput studies that quickly evaluate and screen a group of materials based on their defect behaviour.[27,28,30,37]

Despite progress in accelerating defect predictions, most high-throughput studies are limited in scope. Typically, their training datasets are generated assuming the ideal defect structure inherited from the crystal host, which often lies within a local minimum, thereby trapping a gradient-based optimisation algorithm in a metastable arrangement[38–41]. By yielding incorrect geometries, the predicted defect properties, such as equilibrium concentrations[39,41–43], charge transition levels[39,41,42] and recombination rates[39], are rendered inaccurate[8,44–47]. However, defect structure searching is often too expensive for high-throughput studies that target thousands of defects[30] or materials with complex (defect) energy landscapes, like alloys, disordered solids, and low-symmetry crystals.

In this study, we aim to reduce the computational burden of defect structure searching by introducing a machine-learning surrogate model. We build a dataset containing a set of point defect structures, energies, forces and stresses from first-principles, and use it to fine-tune a universal machine-learning force field (MLFF) and qualitatively explore the energy landscape across 132 defects. Defect reconstructions often follow common motifs[41], especially when comparing similar defects in families of related compounds. By learning the plausible reconstructions undergone by defects in similar hosts, a surrogate model can be used to optimise the initial sampling structures and thus identify the promising, low-energy configurations (Fig. 1), as previously shown for surface adsorbates[48,49] and transition state searches[50].

Results

To assess the ability of a surrogate model to learn defect reconstructions, we will focus on one of the most common — and often strongest in terms of energy-lowering — reconstruction motifs: dimerisation[41,51–73]. Dimers/trimers have been previously reported for numerous vacancies and interstitials, including V_{Se}^{0} in ZnSe, $CuInSe_2$ and $CuGaSe_2$[51], V_S^{0} in ZnS[51], V_{Cd}^{0} in CdTe[39], $V_{Sb}^{0,+1,+2}$ in Sb_2S/Se_3[40,42], $V_{Ti}^{0,-1}$ and V_{Zr}^{0} in $CaZrTi_2O_7$[45], V_{Sb}^{0} in Sb_2O_5[74], O_i^{0} in In_2O_3[55], ZnO[58], Al_2O_3[59], MgO[60,61], CdO[62], SnO_2[63,64],

[1]Thomas Young Centre and Department of Materials, Imperial College London, Exhibition Road, London SW7 2AZ, UK. [2]Thomas Young Centre and Department of Chemistry, Imperial College London, 80 Wood Ln, London W12 7TA, UK. [3]Department of Physics, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. e-mail: a.walsh@imperial.ac.uk

PbO_2[65], CeO_2[66], $BaSnO_3$[75], In_2ZnO_4[67] and $LiNi_{0.5}Mn_{1.5}O_4$[76], Ag_i^0 in AgCl and AgBr[53], V_I^-, I_i^0, Pb_i^0, $Pb^0_{CH_3NH_3}$ and $I^0_{CH_3NH_3}$ in $CH_3NH_3PbI_3$[68–71], Pb_i^0 in $CsPbBr_3$[52], $(CH_3NH_3)_3Pb_2I_7$[54], $(CH_3NH_3)_2Pb(SCN)_2I_2$[72] and Sn_i^0 in $CH_3NH_3SnI_3$[57]. While cation dimerisation has been reported in several hosts (AgCl/Br, $CuInSe_2$, $CuGaSe_2$, ZnS/Se, CdTe, Sb_2S/Se_3, $CH_3NH_3PbI_3$, $CsPbBr_3$, $(CH_3NH_3)_3Pb_2I_7$, $CH_3NH_3SnI_3$)[41,51–54], anion dimers are more common and will be the focus of our study.

To target dimerisation, we consider cation vacancies in low-symmetry metal sulfides/selenides, where their covalent character and soft structures favour dimer formation[41,42,56]. Our first-principles dataset spans 50 hosts (exemplified in Fig. 2a) and 132 neutral cation vacancies, covering 25 elements (Fig. 3b) and 6 space groups. The configurational landscape of each vacancy was explored with the ShakeNBreak method[41,77] by applying 15 chemically-guided distortions to the unperturbed defect structure, followed by geometry optimisation with DFT (see Methods)—resulting in a diverse set of trajectories for each defect and the dataset shown in Fig. 2c.

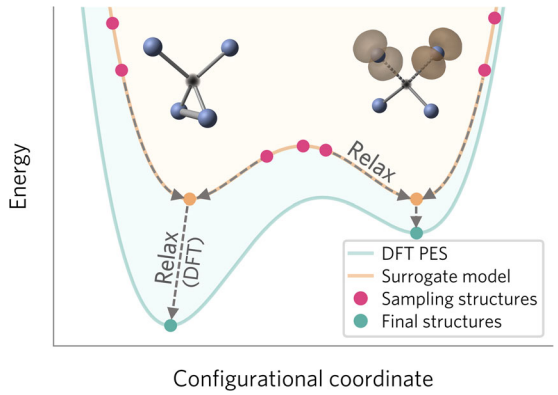


Fig. 1 | Schematic of a machine-learning surrogate model used to accelerate defect structure searching. The computationally efficient model learns the plausible defect reconstructions (local minima in the potential energy surface) and thus reduces the number of candidate structures relaxed with expensive first-principles density functional theory (DFT) calculations.

Defect reconstructions

By analysing our first-principles dataset, we find that 29.9% of the neutral defects undergo symmetry-breaking reconstructions missed by both the standard modelling approach but *also* when applying a rattle distortion (with energy differences between the identified ground state and the relaxed ideal configuration greater than 0.5 eV; Supplementary Table 1, Supplementary Fig. 1). Rattle distortions (i.e. randomised displacements) have been used in recent studies[37] as the prevalence of defect reconstructions have become more recognised. While rattling helps to break the symmetry of the initial defect configuration and escape PES saddle points, it often fails to identify reconstructions with significant energy barriers (i.e. bond formation), highlighting the need for structure searching.

The identified reconstructions are driven by anion–anion bond formation, with the number of new bonds determined by the number of valence electrons lost upon defect formation (Supplementary Fig. 2). In general, energy-lowering structural reconstructions at defects tend to be driven by the localisation of excess charge introduced by the defect formation, through various bonding (re-)arrangements. Here, excess charge refers to the change in valence electrons available for bonding, which is determined by the oxidation state of the original defect atom and the defect charge state—and in fact is the chemical guiding principle used in ShakeNBreak to target likely distortion pathways. For instance, upon forming a neutral antimony vacancy (V_{Sb}^0) in $Sb_2(S/Se)_3$ (where Sb is in the +3 oxidation state), we have removed three bonding electrons and so we have three excess holes. Further changes in the defect charge state will then alter this excess charge (e.g. 2 excess holes in the −1 charge state, or zero excess charge in the 'fully-ionised' -3 charge state). Similarly, for a neutral Li vacancy in Li_4SnS_4, we have removed 1 bonding electron and so we have 1 excess hole (and zero excess charge for the fully ionised -1 state). Analogously for an anion vacancy behaving as a donor defect (as in most semiconductors), it would contribute x excess electrons where -x is the oxidation state of the anion in that compound. Defects resulting in one hole (e.g. V_{Li}^0 in Li_4SnS_4) can easily accommodate the

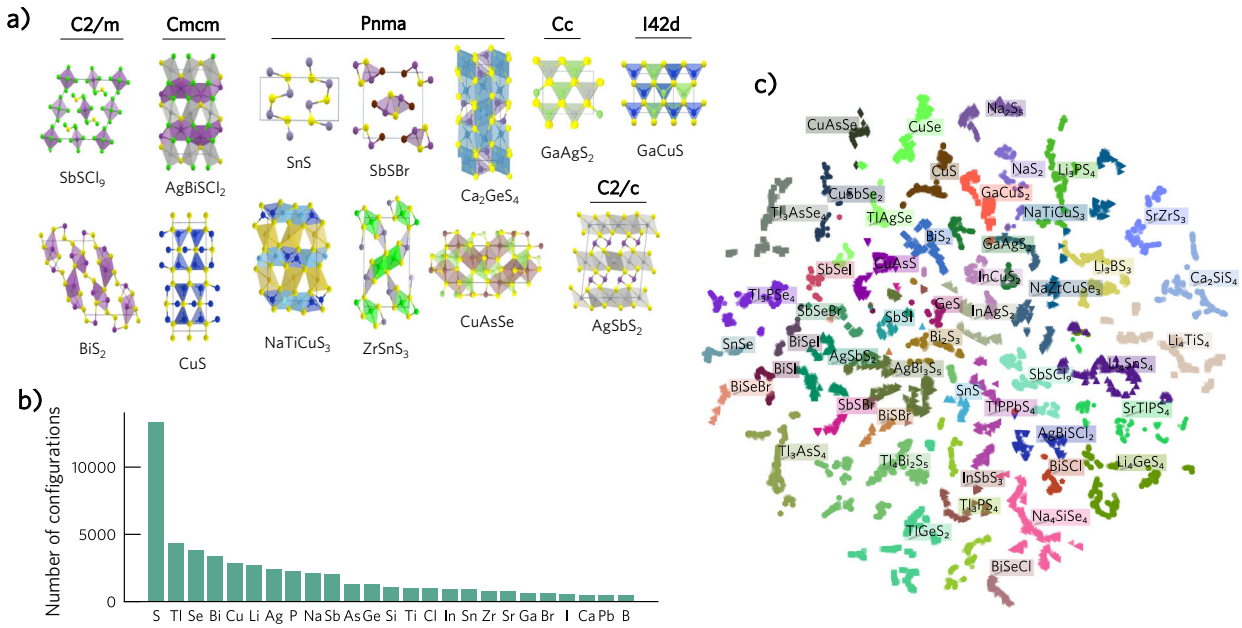


Fig. 2 | Distribution of the defect dataset generated with first-principles calculations. a Example host structures and their respective space groups. b) Number of configurations containing each element. c) Two-dimensional projection of structural similarity for defect configurations. Each configuration is represented with the feature vector generated by the M3GNet model[79,80] (trained on the bulk formation energies of the Materials Project database) and the vector dimensions are reduced using t-distributed stochastic neighbour embedding (t-SNE)[131,132]. The defect configurations are coloured by their host composition (with similar colours indicating compositions with similar MEGNet[133] feature vectors), showing that related chemical systems cluster near each other. For clarity, in (**b**) and (**c**) 10 evenly spaced steps are selected from each relaxation trajectory.

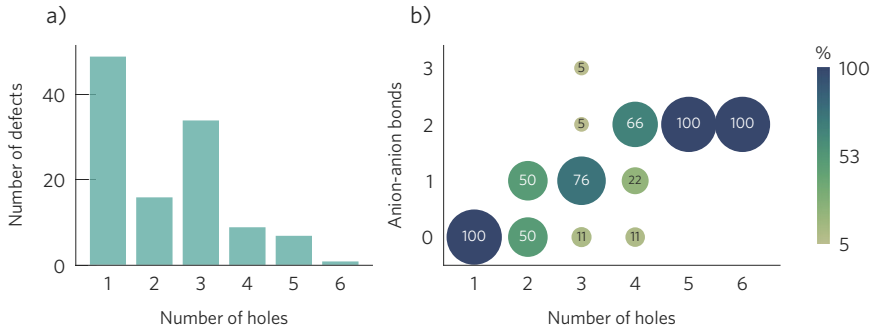


Fig. 3 | Analysis of the point defect dataset. **a** Distribution of number of holes produced upon defect formation. **b** Correlation between the number of anion—anion bonds formed and the number of holes created per defect. The label, colour, and size of the circles indicate the percentage of defects with that number of anion—anion bonds for defects for a given number of holes created.

missing charge without strong reconstructions, while defects with two or more holes (e.g. V_{Bi}^{0} in BiSI) tend to form anion dimers or trimers, as shown in Fig. 3b. As a result, anion–anion bonds are more favourable for more positive defect charge states, and can stabilise unexpected defect oxidation states, as observed previously for V_{Sb}^{+1} in $Sb_2(S/Se)_3$[41,42] and $O_i^{+1,+2}$ in several metal oxides[41,55,58].

There are some exceptions to this trend, where systems are able to accommodate three or more holes without undergoing strong reconstructions. One example is hosts with *d/f* metals that adopt multiple stable oxidation states (e.g. Fe, Co, Cu), which can accommodate a hole by adopting a higher oxidation state[8]. To verify this trend, we compared two isostructural $A^{III}B^{I}S_2$ systems which only differ in the identity of the B cation: V_{Ga}^{0} in Ga**Cu**S_2 and Ga**Ag**S_2; and V_{In}^{0} in In**Cu**S_2 and In**Ag**S_2 (Supplementary Fig. 3). In (Ga/In)**Ag**S_2, two of the holes localise in a S–S bond formed by the vacancy nearest neighbours (NN), while the third hole is split between the remaining two NNs. In contrast, in (Ga/In)**Cu**S_2, no dimer forms since three holes are localised in three of the vacancy NNs and five of the Cu ions closer to the vacancy — with these Cu ions showing shorter Cu–S bonds. The different behaviour of Cu and Ag can be rationalised by considering their second ionisation energies (I2(Cu): 20.3 eV, I2(Ag): 21.5 eV)[78], where the low I2 (and thus higher *d* states) of Cu(I) favours cation oxidation, while the higher I2 of Ag(I) results in a sulfur dimer accommodating two of the holes (Supplementary Fig. 3).

In addition to systems with *d/f* elements, defects with nearby anion–anion bonds can localise the positive charge in these bonds and thus avoid forming new ones. This behaviour is exemplified by $RhSe_2$, where the two symmetry-inequivalent Rh vacancies show different reconstructions. The first vacancy site is surrounded by four Se–Se bonds (Supplementary Fig. 4b), and thus can accommodate the four holes by depopulating the antibonding orbitals of these bonds. In contrast, the second site has only one Se–Se bond neighbouring the vacancy (Supplementary Fig. 4c), and thus has to form an additional Se dimer to accommodate the positive charge.

Beyond chalcogenide dimers, other rearrangements to accommodate positive charge involve chalco-halide (e.g. S–Cl formed by V_{Bi}^{0} in $AgBiSCl_2$) and halide-halide bond formation (e.g. Cl dimers formed by V_{Sb}^{0} in $SbSCl_9$) (Supplementary Fig. 5). Here we note that the zero-dimensional character of $SbSCl_9$ enables this defect to undergo strong distortions forming two Cl dimers (Supplementary Fig. 5). Overall, we highlight the common reconstruction motifs exhibited by different defects in various host structures (Supplementary Fig. 2), facilitating the requisite diversity for a model to learn the plausible reconstructions for a group of related defects.

Model training

To develop a model that can be applied for defect structure searching in *unseen* compositions, we first split our dataset by composition into training, validation and test sets (Supplementary Fig. 6), amounting to 68%, 5% and 27% of defects, respectively. The validation set is then augmented with 5% of the configurations selected for the systems in the training set, to ensure that the structural diversity of the training set is also included for validation (thus evaluating how the model performs for a large diversity of defects and compositions and also how it extrapolates to unseen compositions). This results in training, validation and test sets of 11,955 (63%), 2,100 (11%), and 4830 (26%) configurations, respectively, where configuration denotes a point defect structure with its associated energy, forces and stresses.

To sample the training data, we compared two approaches: (i) a manual method where we sample 10 evenly spaced frames from each relaxation (MS) and (ii) the Dimensionality-Reduced Encoded Clusters approach (DIRECT)[79], which aims to select a robust training set from a complex configurational space. Surprisingly, we find that, when using datasets of similar sizes, the MS approach performs better—with the DIRECT approach only outperforming MS when the final DIRECT dataset is larger than the MS one (Supplementary Table 3). This is because the DIRECT approach mainly samples structures from the initial ionic steps (Supplementary Fig. 9), which correspond to high distortions and thus lead to larger errors for the low energy structures (Supplementary Fig. 10).

As a surrogate model, we aim for a method that takes an initial defect structure and outputs the energy and structure of the locally relaxed configuration. Machine-learning force fields are ideal for this task since they can map regions of the potential energy surface (PES) by learning the energies, forces, and, optionally stresses of a set of training structures. Specifically, we focus on universal graph-based MLFFs, which are trained on relaxation data from diverse databases of bulk crystals[80–83], and thus already incorporate general chemical behaviour. Accordingly, we use a universal MLFF as a base model and fine-tune it with a training set of defect configurations. We have compared different model architectures (M3GNet[80], CHGNet[81] and MACE[84]), elemental reference energies, structure featurisation parameters (graph cutoffs, readout layers) and fine-tuning strategies, which are discussed in detail in the Supporting Information (SI) (Supplementary Notes 1.B). In addition, we compared a model trained on just defect structures, and both defect and bulk structures, with the second case improving performance (Supplementary Table 10 and Supplementary Fig. 12). From these benchmarks, the optimal model architecture and parameters were selected: a M3GNet model[80] with radial and 3-body cutoffs of 5 Å and 4 Å, respectively, and the weighted atom readout function[80,85] (further details in Methods).

Overall, we note that the mean absolute errors for the *absolute* energies in the validation and test sets are significant ($MAE_{E,test}$ = 31.2 meV atom^{-1}, Table 1), but comparable to those obtained in MLFFs used for bulk structure searching of carbon ($MAE_{E,test}$ = 64.8 meV atom^{-1})[86]. However, a more meaningful metric for our purpose is the error for the *relative* energies of each defect configuration relative to its ground state structure ($MAE_{E,rel,test}$ = 11.3 meV atom^{-1}). Further, we mostly care about the low-energy region of the potential energy surface, which can be measured by calculating the relative energy errors for configurations less than ≈ 5 eV above the global minimum, resulting in MAEs of 3.6 meV atom^{-1} ≈ 0.29 eV for an 80 atom supercell.

Beyond these metrics, we calculate the Spearman correlation coefficient (ρ) to measure how well the MLFF and DFT energies are monotonically related (i.e. if greater DFT energies correspond to greater MLFF energies[87]). While the value of ρ for the test set is significantly lower than those obtained with MLFFs developed for *bulk* structure searching for a *single composition* (0.72 versus 0.98–0.999[87]), this was expected considering that our dataset spans a diverse range of compositions and a wide range of energies. While the errors are high, we note that this does not prevent the model from being used as a *qualitative* surrogate of the DFT PES for

structure searching (i.e. identification of local minima), as previously observed for surface adsorbates[88,89].

Model performance

To evaluate the model performance, we apply the trained model to a robust test set, which includes 13 unseen compositions and 32 defects (accounting for 26% and 26.5% of the total number of compositions and defects in our dataset, respectively; Supplementary Fig. 6). For each defect, the MLFF is used to relax the 15 distorted structures generated with ShakeNBreak[77] to sample the defect PES. The MLFF-relaxed structures are then compared to identify the different local minima in the MLFF PES using the SOAP fingerprint[90] of the defect site. These local minima are then further relaxed with DFT. By comparing the ground state identified from the MLFF+DFT approach and full DFT search, we find the former to correctly identify the DFT ground state for 88% of test defects, while simultaneously reducing the number of DFT calculations required by 73% (Table 2) and accelerating structure searching by a factor of 13 (Supplementary Notes 1.C4). In addition, it identifies a more favourable structure than the ones found in the DFT search for $V_{Ge,9}$ in $TlGeS_2$, with an energy lowering of 0.5 eV (Supplementary Fig. 15). The 12% of failed cases, where the MLFF ground state structure differed from the DFT one, mostly involve complex hosts. For instance, V_{Sn} in Li_4SnS_4 has a complex DFT energy surface, which traps most of the relaxations in very high energy basins (Supplementary Fig. 19). PESs of similar complexity are displayed by the iso-structural systems that were included in the training set (Li_4GeS_4 and Li_4TiS_4; Supplementary Fig. 19), which biases our training data to the high energy region of the PES for these compositions and thus hinders learning the low-energy region. Accordingly, the training data for these systems can be improved by reducing the magnitude of the distortion used by ShakeNBreak to generate their sampling structures; which would improve model performance. Other defects for which the surrogate model misses the most stable structure are $V_{Tl,0}$ in $TlGeS_2$ and V_{Bi} in BiSeI — yet in both cases the DFT and MLFF+DFT structures are very similar and differ by small energy differences (0.1 and 0.2 eV, respectively) (Supplementary Figures 16 and 17). In all failed cases, while the model misses the full DFT ground state, it still correctly predicts a favourable reconstruction, that lowers the energy compared to the relaxed ideal configuration.

Beyond identifying the correct ground state in the majority of cases, the model has indirectly learned the correlation between the number of holes and the number of formed dimers. For defects with 1 missing electron, the candidate structures generated by the surrogate model rarely contained anion–anion bonds; while for defects with more missing electrons, the model often identifies at least one local minima with a dimer.

The decreased performance observed for out-of-sample compositions less similar to the training set posed the question of what performance could be achieved if targeting a family of more related systems. To consider more similar host compositions, we select the chalcohalide systems from our dataset and split them composition-wise into training, validation and test sets as described in Supplementary Notes 1.C5. After training the model on the training set and applying it to the unseen test defects (details in Supplementary Notes 1.C5), we find that the model identifies the correct ground state for all test cases, and achieves lower mean absolute errors compared to the full model. This suggests that higher accuracy can be achieved when targeting more similar host structures, which is likely the case in most high-throughput defect studies.

Our current trained model is limited to neutral cation vacancies in metal sulphides/selenides. However, the approach can be extended to a different compositional space or defect type by first generating a custom training set through first-principles calculations and using it to fine-tune the universal bulk MLFF.

Application to alloys

Beyond high-throughput studies of many single-phase materials, the surrogate model can also be used to accelerate structure searching in alloys or disordered solids, which is computationally challenging due to the high number of local host compositions and inequivalent defects to consider[91]. The distinct local or site environments of a given defect can significantly affect its properties[34,35,92–103], altering formation and migration energies by up to 1.5 eV[34,35,93,100–103]. Properly sampling various site environments is key to characterise the defect behaviour in such cases.

We consider the case of cadmium vacancies in the $CdSe_xTe_{(1-x)}$ ($x = 0, 0.2, 0.3, 0.5, 0.6, 0.8, 0.9, 1$) pseudo-binary alloy. For each composition, a supercell is generated through random substitution of Te sites, and the Cd sites with a unique nearest neighbour chemical environment are considered (e.g. Cd surrounded by 4 Te; by 3 Te and 1 Se; by 2 Se and 2 Te, etc) (Fig. 4a). The configurational landscape of each vacancy is explored with the ShakeNBreak method (14 sampling structures), using the relaxations from the pure compositions as the training and validation data while the mixed systems ($0 < x < 1$) are reserved as the test set.

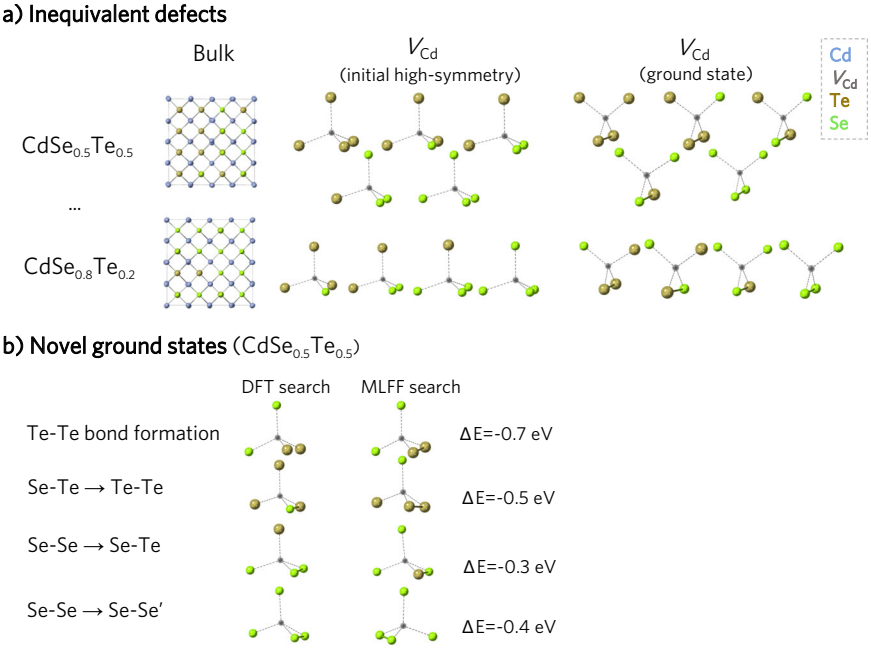


Fig. 4 | Structures for V_{Cd} in the $CdSe_xTe_{1-x}$ alloys. a Inequivalent defect environments for two of the $CdSe_xTe_{1-x}$ alloys ($x = 0.5, 0.8$). **b** Examples of the ground state configurations *only* identified through the finer MLFF+DFT search. These reconstructions are driven by either forming a dimer (e.g. Te–Te bond formation), forming a more favourable anion–anion bond (Te–Te instead of Se–Te; Se–Te instead of Se–Se) or forming the same type of anion bond (Se-Se) but breaking weaker anion-cation bonds between the defect nearest neighbours and the defect next nearest neighbours.

After fine-tuning the surrogate model (MLFF) on the training configurations (details in Methods), it is applied to all alloys to perform the structure searching calculations, allowing a more extensive sampling than for the DFT search (31 sampling structures). From the MLFF-relaxed structures, the unique configurations are selected for further relaxation with DFT and compared with the results from the DFT-only search. This comparison shows that the model successfully identifies the ground state for all defects, even in cases where the defects form Te-Se bonds not seen in the training set – which only included the Te-Te and Se-Se bonds formed by V_{Cd}(CdTe) and V_{Cd}(CdSe), respectively. Although Te-Se bonds were not present in our defect training set, they were included in the Materials Project (bulk) training set, thus suggesting the benefit of transfer learning for model generalisability.

More significantly, for 70% of the defects, the model identifies a more favourable ground state missed in the coarser DFT search (with a mean energy lowering of −0.4 eV, Supplementary Notes 1.E). These reconstructions are driven by forming a more favourable anion–anion bond (Fig. 4b) and missed in the DFT-only search due to the coarser sampling performed. This illustrates the advantage of the faster surrogate model to tackle defects with complex configurational landscapes, that require a more exhaustive exploration than a DFT-based search would allow, like alloys, compositionally disordered materials[104–107], and low-symmetry or multinary systems with many degrees of freedom in their PES.

Discussion

By building a dataset for defect structure searching, we have demonstrated the prevalence of defect reconstructions missed by the standard modelling approach – and thus the need to perform structure searching in high-throughput defect studies. To reduce the associated computational burden, we have developed a surrogate model by fine-tuning a universal machine-learning force field on defect configurations. By qualitatively learning the defect configurational landscapes, the trained model successfully predicts low-energy defect structures for *unseen* defect environments in *unseen* compositions, thereby reducing the number of DFT calculations by 73%. While our current model is limited to neutral cation vacancies in metal chalcogenides, the methodology can be applied to different defect types or compositional spaces. In addition, our openly-available dataset could be used to measure the out-of-distribution performance of universal MLFFs[108] by testing the ability to extrapolate from learned bulk motifs to defect environments.

Beyond accelerating structure searching in high-throughput studies, this approach is ideal for systems with a complex defect landscape, like alloys, disordered, or low-symmetry materials where their many inequivalent defects make it intractable to explicitly calculate all of them with accurate DFT methods. By using a surrogate model, we can consider a range of alloy compositions and all inequivalent defects, while performing a more exhaustive sampling of the PES — thereby identifying more favourable reconstructions missed in the (coarser) DFT-based search. Beyond (pseudo-)binary alloys, this approach could be extended to model more chemically complex systems, like high-entropy alloys, where the MLFF could be trained on defects of the constituent binary systems and applied to the ternary, quaternary, or high-entropy alloys.

A current limitation of this strategy is the handling of defects in distinct charge states, which have different energy landscapes and structural configurations (e.g. a defect in two different charge states can have a common local structure with different energies). The approach could handle the potential energy landscape for each charge state independently (e.g. training a separate model for defects in the -1 charge state). To consider different charge states simultaneously, the net charge state can be encoded as a graph global attribute[109]. However, a more descriptive encoding could be achieved by using fourth-generation MLFFs that include atomic charges[110]. Beyond accounting for the defect charge, another improvement could be MLFFs that are fine-tuned on-the-fly during geometry optimisations. As shown for surface absorbates[88,89], this strategy would accelerate the defect geometry optimisation by skipping many ionic steps that are performed with the surrogate model. Overall, we note the promise of surrogate models to accelerate and increase the accuracy of defect modelling, whether this is by improving structure searching, accounting for metastable configurations[111,112], enabling the calculation of defect formation entropies[109,112], accelerating defect migration studies[113] or going beyond the dilute limit[107].

Methods

High-throughput vacancies in chalcogenide hosts

The conventional supercell approach for modelling defects in periodic solids was used[114]. To reduce periodic image interactions, supercell dimensions of at least 10 Å in each direction[115] were employed. To explore the configurational landscape of each defect, we used the ShakeNBreak code[77], with a distortion increment of 0.1 and the default rattle standard deviation (10% of the nearest neighbour distance in the bulk supercell). This strategy results in 14 sampling structures. In addition to these, to ensure that dimerisation was properly sampled, we also generated a sampling structure where two defect neighbours were pushed towards each other with a separation of 2 Å, resulting in a total of 15 initial configurations. Due to the limitation of universal MLFFs to describe charge, we only considered one charge state of the defects. We chose the neutral state for several reasons. First, it is often stable for cation vacancies in metal chalcogenides. Secondly, it is usually included within the potential charge states to be calculated for a given defect (e.g. generally ranging from the fully ionised state to (at least) the neutral one), and it often has a complex potential energy landscape. However, we note that we did not check whether it was the thermodynamically stable state for each defect.

All reference calculations were performed with Density Functional Theory using the exchange-correlation functional HSE06[116] and the projector augmented wave method[117], as implemented in the Vienna Ab initio Simulation Package[118,119]. Calculations for the pristine unit cells were performed using a plane wave energy cutoff of 585 eV and sampling reciprocal space with a Monkhorst-Pack mesh of density 900 *k*-points/site. The convergence thresholds for the geometry optimisations were set to 10^{-6} eV and 10^{-5}eV Å^{-1} for energy and forces, respectively. Defect relaxations were performed with the Γ-point approximation, which is accurate enough for defect structure searching[41], and with a plane wave energy cutoff of 350 eV. The energy and force thresholds for defect relaxations were set to 10^{-4} eV and 10^{-2}eV Å^{-1}, respectively. We note that these settings were selected for an efficient exploration of the defect configurational landscape due to the high number of relaxations required for structure searching. In a full defect study aiming for high accuracy, once the ground state configuration is identified with these settings, it should be further relaxed with tighter convergence thresholds and account for spin-orbit coupling when necessary (elements from period five/six and below).

To automate the generation of input files, we designed a workflow using aiida[120–122], pymatgen[123–125], pymatgen-analysis-defects[126,127], ASE[128], doped[129] and ShakeNBreak[77]. This code is available from https://github.com/ireaml/defects_workflow.git. The datasets and trained models are available from the Zenodo repository with https://doi.org/10.5281/zenodo.10579527.

To generate the training and test set for the machine learning model, we processed the DFT defect relaxation data by removing unreasonably high-energy configurations (e.g. structures with positive energies), as they decreased model performance. After cleaning the data, 10 evenly-spaced ionic steps were selected from each relaxation. We used the M3GNet model[80], as implemented in ref.85, with radial and 3 body cutoffs of 5 Å and 4 Å, respectively, and the weighted atom readout function. The loss function was defined as a combination of the mean squared errors for the energies, forces and stresses, with respective weights of 1, 1 and 0.1[80,85]. For fine-tuning, the model was initialised with the weights from the trained bulk crystal model[85] and then trained on the defect training set (see Supplementary Notes 1.B5 for further details). A batch size of 4 and an exponential learning rate scheduler with an initial rate of 5×10^{-4} were used. The model was trained on a Quadro RTX6000

GPU until the validation errors were converged (30 epochs, 5.3 hours) (Supplementary Fig. 13).

MLFF geometry optimisation was performed with the FIRE algorithm[130], as implemented in the ASE package[128], until the mean force was lower than 10^{-5}eV Å^{-1} or the number of ionic steps exceeded 1500, which were found to be reasonable thresholds. After relaxing the sampling structures with the model, we identified the different local minima or configurations by calculating the cosine distance between the SOAP descriptor[90] for the defect site of each configuration, which was found to be an effective metric for identifying different defect motifs. We note that using the SOAP fingerprint of the defect site was more robust than considering the energies or the root mean squared displacement between the structures. The first case can miss local minima if these have similar energies in the MLFF PES, while the second was more sensitive to structural differences far from the defect site. The parameters used to generate the SOAP descriptor were: $r = 5$ Å, $n_{max} = 10$, $l_{max} = 10$, $\sigma = 1.0$ Å, for the local cutoff, number of radial basis functions, maximum degree of spherical harmonics, and the standard deviation of the Gaussian functions used to expand the atomic density, respectively. To evaluate the correlation between DFT and MLFF energies, the Spearman coefficient was calculated for each defect *independently*, and then averaged across defects.

Application to the $CdSe_xTe_{(1-x)}$ alloy

To generate the supercells for the mixed compositions in $CdSe_xTe_{(1-x)}$ ($x = 0.2, 0.3, 0.5, 0.6, 0.8, 0.9$), we used random substitution of Te sites with Se. For each supercell, we consider the Cd sites with a unique nearest neighbour chemical environment as vacancy sites (e.g. Cd surrounded by 4 Te; by 3 Te and 1 Se; by 2 Se and 2 Te etc), and generate the vacancy high-symmetry structures with pymatgen[123–125]. For the DFT-based exploration of the PES, we apply ShakeNBreak with default parameters, generating 14 sampling structures, which were relaxed with DFT as previously described.

To generate the dataset, we again processed the defect relaxation data by removing unreasonably high-energy configurations (>15 eV above the defect ground state configuration). As the training set, we used a combination of defect and bulk configurations: 45 evenly-spaced frames from the 14 relaxations of V_{Cd} in CdTe and CdSe, and 20 frames from the relaxations of each pristine system, resulting in a total of 1420 frames. For validation, we selected 5 unseen configurations from the 14 relaxations of V_{Cd} in CdTe and CdSe (total of 140 frames). The M3GNet surrogate model was trained with similar parameters as previously described and until the errors were converged (80 epochs, see Supplementary Fig. 26).

To perform a finer exploration of the PES with the surrogate model, we applied a set of bond distortions generated by ShakeNBreak (−0.6, −0.5, −0.4, −0.3, −0.2) to all unique pair combinations of nearest neighbours (e.g. for a V_{Cd} surrounded by two Te and two Se anions, Te(1), Te(2), Se(1) and Se(2), we considered the pairs Te(1)-Te(2), Te(1)-Se(1), Te(1)-Se(2), Te(2)-Se(1), Te(2)-Se(2), and Se(1)-Se(2)). By default, for a defect with two missing electrons like V_{Cd}^{0}, ShakeNBreak only applies the bond distortions to the two atoms closest to the defect. This is typically a reliable approach for most *pure* systems, but can miss reconstructions for alloys with complex defect environments (e.g. V_{Cd} surrounded by a mix of Te and Se anions). The model application and analysis were performed as described in the previous section.

Data availability

The datasets and trained models are available from the Zenodo repository with https://doi.org/10.5281/zenodo.10579527.

Code availability

The code used to generate the defect dataset is available from https://github.com/ireaml/defects_workflow.git.

Received: 19 January 2024; Accepted: 24 May 2024;
Published online: 06 June 2024

References

1. Sambur, J. & Brgoch, J. Unveiling the hidden influence of defects via experiment and data science. *Chem. Mater.* **35**, 7351–7354 (2023).
2. Shockley, W. & Read, W. T. Statistics of the recombinations of holes and electrons. *Phys. Rev.* **87**, 835–842 (1952).
3. Kim, S., Márquez, J. A., Unold, T. & Walsh, A. Upper limit to the photovoltaic efficiency of imperfect crystals from first principles. *Energy Environ. Sci.* **13**, 1481–1491 (2020).
4. Maier, J. Thermodynamics of electrochemical lithium storage. *Angew. Chem. Int. Ed.* **52**, 4998–5026 (2013).
5. Squires, A. G. et al. Low electronic conductivity of $Li_7La_3Zr_2O_{12}$ solid electrolytes from first principles. *Phys. Rev. Mater.* **6**, 085401 (2022).
6. Li, W. et al. Defect engineering for fuel cell electrocatalysts. *Adv. Mater.* **32**, 1907879 (2020).
7. Pastor, E. et al. Electronic defects in metal oxide photocatalysts. *Nat. Rev. Mater.* **7**, 503–521 (2022).
8. Kehoe, A. B., Scanlon, D. O. & Watson, G. W. Role of lattice distortions in the oxygen storage capacity of divalently doped CeO_2. *Chem. Mater.* **23**, 4464–4468 (2011).
9. Ivády, V., Abrikosov, I. A. & Gali, A. First principles calculation of spin-related quantities for point defect qubit research. *npj Comput Mater.* **4**, 76 (2018).
10. Weber, J. R. et al. Quantum computing with defects. *Proc. Natl Acad. Sci. USA* **107**, 8513–8518 (2010).
11. Thomas, J. et al. A substitutional quantum defect in WS_2 discovered by high-throughput computational screening and fabricated by site-selective STM manipulation. *Nat. Commun*. **15**, 3556 (2024).
12. Dreyer, C. E., Alkauskas, A., Lyons, J. L., Janotti, A. & Van de Walle, C. G. First-principles calculations of point defects for quantum technologies. *Annu. Rev. Mater. Res.* **48**, 1–26 (2018).
13. Yan, Q., Kar, S., Chowdhury, S. & Bansil, A. The case for a defect genome initiative. *Adv. Mater.* **36**, 2303098 (2024).
14. Davidsson, J., Bertoldo, F., Thygesen, K. S. & Armiento, R. Absorption versus adsorption: in 2D materials. *npj 2D Mater. Appl.* **7**, 26 (2023).
15. Sluydts, M., Pieters, M., Vanhellemont, J., Speybroeck, V. V. & Cottenier, S. High-throughput screening of extrinsic point defect properties in Si and Ge: database and applications. *Chem. Mater.* **29**, 975–984 (2016).
16. Bertoldo, F., Ali, S., Manti, S. & Thygesen, K. S. Quantum point defects in 2d materials—the QPOD database. *npj Comput. Mater.* **8**, 56 (2022).
17. Huang, P. et al. Unveiling the complex structure-property correlation of defects in 2D materials based on high throughput datasets. *npj 2D Mater. Appl* **7**, 1–10 (2023).
18. Medasani, B. et al. Predicting defect behavior in B2 intermetallics by merging ab initio modeling and machine learning. *npj Comput Mater.* **2**, 1–10 (2016).
19. Rahman, M. H. et al. Accelerating defect predictions in semiconductors using graph neural networks. *APL Mach. Learn.* **2**, 016122 (2024).
20. Ivanov, V. et al. Database of semiconductor point-defect properties for applications in quantum technologies. Preprint at https://arxiv.org/abs/2303.16283 (2023).
21. Kumagai, Y., Tsunoda, N., Takahashi, A. & Oba, F. Insights into oxygen vacancies from high-throughput first-principles calculations. *Phys. Rev. Mater.* **5**, 123803 (2021).
22. Deml, A. M., Holder, A. M., O'Hayre, R. P., Musgrave, C. B. & Stevanović, V. Intrinsic material properties dictating oxygen vacancy formation energetics in metal oxides. *J. Phys. Chem. Lett.* **6**, 1948–1953 (2015).
23. Broberg, D. et al. High-throughput calculations of charged point defect properties with semi-local density functional theory—performance benchmarks for materials screening applications. *npj Comput. Mater.* **9**, 72 (2023).

24. Mannodi-Kanakkithodi, A. et al. Universal machine learning framework for defect predictions in zinc blende semiconductors. *Patterns* **3**, 100450 (2022).
25. Varley, J. B., Samanta, A. & Lordi, V. Descriptor-based approach for the prediction of cation vacancy formation energies and transition levels. *J. Phys. Chem. Lett.* **8**, 5059–5063 (2017).
26. Wan, Z., Wang, Q.-D., Liu, D. & Liang, J. Data-driven machine learning model for the prediction of oxygen vacancy formation energy of metal oxide materials. *Phys. Chem. Chem. Phys.* **23**, 15675–15684 (2021).
27. Wexler, R. B., Gautam, G. S., Stechel, E. B. & Carter, E. A. Factors governing oxygen vacancy formation in oxide perovskites. *J. Am. Chem. Soc.* **143**, 13212–13227 (2021).
28. Frey, N. C., Akinwande, D., Jariwala, D. & Shenoy, V. B. Machine learning-enabled design of point defects in 2d materials for quantum and neuromorphic information processing. *ACS Nano* **14**, 13406–13417 (2020).
29. Sharma, V., Kumar, P., Dev, P. & Pilania, G. Machine learning substitutional defect formation energies in ABO_3 perovskites. *J. Appl. Phys.* **128**, 034902 (2020).
30. Baldassarri, B. et al. Oxygen vacancy formation energy in metal oxides: High-throughput computational studies and machine-learning predictions. *Chem. Mater.* **35**, 10619–10634 (2023).
31. Park, S. et al. Exploring the latent chemical space of oxygen vacancy formation energy by a machine learning ensemble. *ACS Mater. Lett.* **6**, 66–72 (2024).
32. Kazeev, N. et al. Sparse representation for machine learning the properties of defects in 2D materials. *npj Comput. Mater.* **9**, 113 (2023).
33. Choudhary, K. & Sumpter, B. G. Can a deep-learning model make fast predictions of vacancy formation in diverse materials? *AIP Adv.* **13**, 095109 (2023).
34. Zhao, X., Yu, S., Zheng, J., Reece, M. J. & Zhang, R.-Z. Machine learning of carbon vacancy formation energy in high-entropy carbides. *J. Eur. Ceram. Soc.* **43**, 1315–1321 (2023).
35. Manzoor, A. et al. Machine learning based methodology to predict point defect energies in multi-principal element alloys. *Front. Mater.* **8**, 673574 (2021).
36. Polak, M. P., Jacobs, R., Mannodi-Kanakkithodi, A., Chan, M. K. Y. & Morgan, D. Machine learning for impurity charge-state transition levels in semiconductors from elemental properties using multi-fidelity datasets. *J. Chem. Phys.* **156**, 114110 (2022).
37. Witman, M. D., Goyal, A., Ogitsu, T., McDaniel, A. H. & Lany, S. Defect graph neural networks for materials discovery in high-temperature clean-energy applications. *Nat. Comput. Sci.* **3**, 675–686 (2023).
38. Arrigoni, M. & Madsen, G. K. H. Evolutionary computing and machine learning for discovering of low-energy defect configurations. *npj Comput. Mater.* **7**, 1–13 (2021).
39. Kavanagh, S. R., Walsh, A. & Scanlon, D. O. Rapid recombination by cadmium vacancies in CdTe. *ACS Energy Lett.* **6**, 1392–1398 (2021).
40. Mosquera-Lois, I. & Kavanagh, S. R. In search of hidden defects. *Matter* **4**, 2602–2605 (2021).
41. Mosquera-Lois, I., Kavanagh, S. R., Walsh, A. & Scanlon, D. O. Identifying the ground state structures of point defects in solids. *npj Comput. Mater.* **9**, 1–11 (2023).
42. Wang, X., Kavanagh, S. R., Scanlon, D. O. & Walsh, A. Four-electron negative-*U* vacancy defects in antimony selenide. *Phys. Rev. B* **108**, 134102 (2023).
43. Wang, X., Kavanagh, S. R., Scanlon, D. O. & Walsh, A. Upper efficiency limit of Sb_2Se_3 solar cells. *Joule* **8**, 1–18 (2024).
44. Morris, A. J., Pickard, C. J. & Needs, R. J. Hydrogen/nitrogen/oxygen defect complexes in silicon from computational searches. *Phys. Rev. B* **80**, 144112 (2009).
45. Mulroue, J., Morris, A. J. & Duffy, D. M. Ab initio study of intrinsic defects in zirconolite. *Phys. Rev. B* **84**, 094118 (2011).
46. Al-Mushadani, O. K. & Needs, R. J. Free-energy calculations of intrinsic point defects in silicon. *Phys. Rev. B* **68**, 235205 (2003).
47. Kononov, A., Lee, C.-W., Shapera, E. & Schleife, A. Identifying native point defect configurations in α-alumina. *J. Phys. Condens.* **35**, 334002 (2023).
48. Schaarschmidt, M. et al. Learned force fields are ready for ground state catalyst discovery. Preprint at https://arxiv.org/abs/2209.12466 (2022).
49. Lan, J. et al. AdsorbML: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials. *npj Comput. Mater.* **9**, 172 (2023).
50. Heinen, S., von Rudorff, G. F. & von Lilienfeld, O. A. Transition state search and geometry relaxation throughout chemical compound space with quantum machine learning. *J. Chem. Phys.* **157**, 221102 (2022).
51. Lany, S. & Zunger, A. Metal-dimer atomic reconstruction leading to deep donor states of the anion vacancy in II-VI and chalcopyrite semiconductors. *Phys. Rev. Lett.* **93**, 156404 (2004).
52. Kang, J. & Wang, L.-W. High defect tolerance in lead halide perovskite $CsPbBr_3$. *J. Phys. Chem. Lett.* **8**, 489–493 (2017).
53. Wilson, D. J., Sokol, A. A., French, S. A. & Catlow, C. R. A. Defect structures in the silver halides. *Phys. Rev. B* **77**, 064115 (2008).
54. Zhao, Y. et al. Correlations between immobilizing ions and suppressing hysteresis in perovskite solar cells. *ACS Energy Lett.* **1**, 266–272 (2016).
55. Ágoston, P., Erhart, P., Klein, A. & Albe, K. Geometry, electronic structure and thermodynamic stability of intrinsic point defects in indium oxide. *J. Phys. Condens. Matter* **21**, 455801 (2009).
56. Han, D., Du, M.-H., Dai, C.-M., Sun, D. & Chen, S. Influence of defects and dopants on the photovoltaic performance of Bi_2S_3: first-principles insights. *J. Mater. Chem. A* **5**, 6200–6210 (2017).
57. Meggiolaro, D., Ricciarelli, D., Alasmari, A. A., Alasmary, F. A. S. & De Angelis, F. Tin versus lead redox chemistry modulates charge trapping and self-doping in tin/lead iodide perovskites. *J. Phys. Chem. Lett.* **11**, 3546–3556 (2020).
58. Erhart, P., Klein, A. & Albe, K. First-principles study of the structure and stability of oxygen defects in zinc oxide. *Phys. Rev. B* **72**, 085213 (2005).
59. Sokol, A. A., Walsh, A. & Catlow, C. R. A. Oxygen interstitial structures in close-packed metal oxides. *Chem. Phys. Lett.* **492**, 44–48 (2010).
60. Evarestov, R. A., Jacobs, P. W. M. & Leko, A. V. Oxygen interstitials in magnesium oxide: a band-model study. *Phys. Rev. B* **54**, 8969–8972 (1996).
61. Kotomin, E. A. & Popov, A. I. Radiation-induced point defects in simple oxides. *Nucl. Instrum. Methods Phys. Res. B* **141**, 1–15 (1998).
62. Burbano, M., Scanlon, D. O. & Watson, G. W. Sources of conductivity and doping limits in CdO from hybrid density functional theory. *J. Am. Chem. Soc.* **133**, 15065–15072 (2011).
63. Scanlon, D. O. & Watson, G. W. On the possibility of p-type SnO_2. *J. Mater. Chem.* **22**, 25236–25245 (2012).
64. Godinho, K. G., Walsh, A. & Watson, G. W. Energetic and electronic structure analysis of intrinsic defects in SnO_2. *J. Phys. Chem. C* **113**, 439–448 (2009).
65. Scanlon, D. O. et al. Nature of the band gap and origin of the conductivity of PbO_2 revealed by theory and experiment. *Phys. Rev. Lett.* **107**, 246402 (2011).
66. Keating, P. R. L., Scanlon, D. O., Morgan, B. J., Galea, N. M. & Watson, G. W. Analysis of intrinsic defects in CeO_2 using a Koopmans-like GGA +U approach. *J. Phys. Chem. C.* **116**, 2443–2452 (2012).
67. Walsh, A., Da Silva, J. L. F. & Wei, S.-H. Interplay between order and disorder in the high performance of amorphous transparent conducting oxides. *Chem. Mater.* **21**, 5119–5124 (2009).
68. Whalley, L. D., Crespo-Otero, R. & Walsh, A. H-center and V-center defects in hybrid halide perovskites. *ACS Energy Lett.* **2**, 2713–2714 (2017).
69. Agiorgousis, M. L., Sun, Y.-Y., Zeng, H. & Zhang, S. Strong covalency-induced recombination centers in perovskite solar cell material $CH_3NH_3PbI_3$. *J. Am. Chem. Soc.* **136**, 14570–14575 (2014).

70. Whalley, L. D. et al. Giant Huang-Rhys factor for electron capture by the iodine intersitial in perovskite solar cells. *J. Am. Chem. Soc.* **143**, 9123–9128 (2021).
71. Motti, S. G. et al. Defect activity in lead halide perovskites. *Adv. Mater.* **31**, 1901183 (2019).
72. Xiao, Z., Meng, W., Wang, J. & Yan, Y. Defect properties of the two-dimensional $(CH_3NH_3)_2Pb(SCN)_2I_2$ perovskite: a density-functional theory study. *Phys. Chem. Chem. Phys.* **18**, 25786–25790 (2016).
73. Na-Phattalung, S. et al. First-principles study of native defects in anatase TiO_2. *Phys. Rev. B* **73**, 125205 (2006).
74. Li, K., Willis, J., Kavanagh, S. R. & Scanlon, D. O. Computational prediction of an antimony-based n-type transparent conducting oxide: F-doped Sb_2O_5. *Chem. Mater.* **36**, 2907–2916 (2024).
75. Scanlon, D. O. Defect engineering of $basno_3$ for high-performance transparent conducting oxide applications. *Phys. Rev. B* **87**, 161201 (2013).
76. Cen, J., Zhu, B., Kavanagh, S. R., Squires, A. G. & Scanlon, D. O. Cation disorder dominates the defect chemistry of high-voltage $LiMn_{1.5}Ni_{0.5}O_4$ (LMNO) spinel cathodes. *J. Mater. Chem. A* **11**, 13353–13370 (2023).
77. Mosquera-Lois, I., Kavanagh, S. R., Walsh, A. & Scanlon, D. O. ShakeNBreak: navigating the defect configurational landscape. *J. Open Source Softw.* **7**, 4817 (2022).
78. NIST Chemistry WebBook. https://doi.org/10.18434/M32147 (Accessed May 2023).
79. Qi, J., Ko, T. W., Wood, B. C., Pham, T. A. & Ong, S. P. Robust training of machine learning interatomic potentials with dimensionality reduction and stratified sampling. *npj Comput. Mater.* **10**, 1–11 (2024).
80. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728 (2022).
81. Deng, B. et al. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Mach. Intell.* **5**, 1031–1041 (2023).
82. Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).
83. Batatia, I. et al. A foundation model for atomistic materials chemistry. Preprint at https://arxiv.org/abs/2401.00096 (2024).
84. Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csanyi, G. MACE: higher order equivariant message passing neural networks for fast and accurate force fields. *Adv. Neural Inf. Process Syst.* **35**, 11423–11436 (2022).
85. Chen, C. & Ong, S. P. M3GNet (version 0.2.4). GitHub https://github.com/materialsvirtuallab/m3gnet (2023).
86. Salzbrenner, P. T. et al. Developments and further applications of ephemeral data derived potentials. *J. Chem. Phys.* **159**, 144801 (2023).
87. Pickard, C. J. Ephemeral data derived potentials for random structure search. *Phys. Rev. B* **106**, 014102 (2022).
88. Musielewicz, J., Wang, X., Tian, T. & Ulissi, Z. FINETUNA: fine-tuning accelerated molecular simulations. *Mach. Learn. Technol.* **3**, 03LT01 (2022).
89. Jung, H., Sauerland, L., Stocker, S., Reuter, K. & Margraf, J. T. Machine-learning driven global optimization of surface adsorbate geometries. *npj Comput Mater.* **9**, 114 (2023).
90. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
91. Hu, Y.-J. First-principles approaches and models for crystal defect energetics in metallic alloys. *Comput. Mater. Sci.* **216**, 111831 (2023).
92. Piochaud, J. B. et al. First-principles study of point defects in an fcc Fe-10Ni-20Cr model alloy. *Phys. Rev. B* **89**, 024101 (2014).
93. Guan, H. et al. Chemical environment and magnetic moment effects on point defect formations in CoCrNi-based concentrated solid-solution alloys. *Acta Mater.* **187**, 122–134 (2020).
94. Rio, E. D. et al. Formation energy of vacancies in FeCr alloys: dependence on Cr concentration. *J. Nucl. Mater.* **408**, 18–24 (2011).
95. Zhang, Y. et al. Influence of chemical disorder on energy dissipation and defect evolution in concentrated solid solution alloys. *Nat. Commun.* **6**, 8736 (2015).
96. Zhang, Y. et al. Atomic-level heterogeneity and defect dynamics in concentrated solid-solution alloys. *Curr. Opin. Solid State Mater. Sci.* **21**, 221–237 (2017).
97. Arora, G., Bonny, G., Castin, N. & Aidhy, D. S. Effect of different point-defect energetics in $Ni_{80}X_{20}$ (X=Fe, Pd) on contrasting vacancy cluster formation from atomistic simulations. *Acta Mater.* **15**, 100974 (2021).
98. Zhao, S., Stocks, G. M. & Zhang, Y. Defect energetics of concentrated solid-solution alloys from ab initio calculations: $Ni_{0.5}Co_{0.5}$, $Ni_{0.5}Fe_{0.5}$, $Ni_{0.8}Fe_{0.2}$ and $Ni_{0.8}Cr_{0.2}$. *Phys. Chem. Chem. Phys.* **18**, 24043–24056 (2016).
99. Manzoor, A. & Zhang, Y. Influence of defect thermodynamics on self-diffusion in complex concentrated alloys with chemical ordering. *JOM* **74**, 4107–4120 (2022).
100. Zhao, S., Egami, T., Stocks, G. M. & Zhang, Y. Effect of d electrons on defect properties in equiatomic NiCoCr and NiCoFeCr concentrated solid solution alloys. *Phys. Rev. Mater.* **2**, 013602 (2018).
101. Li, C. et al. First principle study of magnetism and vacancy energetics in a near equimolar NiFeMnCr high entropy alloy. *J. Appl. Phys.* **125**, 155103 (2019).
102. Manzoor, A., Zhang, Y. & Aidhy, D. S. Factors affecting the vacancy formation energy in $Fe_{70}Ni_{10}Cr_{20}$ random concentrated alloy. *Comput. Mater. Sci.* **198**, 110669 (2021).
103. Muzyk, M., Nguyen-Manh, D., Kurzydłowski, K. J., Baluc, N. L. & Dudarev, S. L. Phase stability, point defects, and elastic properties of W-V and W-Ta alloys. *Phys. Rev. B* **84**, 104115 (2011).
104. Wang, Y. et al. Cation disorder engineering yields $AgBiS_2$ nanocrystals with enhanced optical absorption for efficient ultrathin solar cells. *Nat. Photon.* **16**, 235–241 (2022).
105. Williford, R., Weber, W., Devanathan, R. & Gale, J. Effects of cation disorder on oxygen vacancy migration in $Gd_2Ti_2O_7$. *J. Electroceram.* **3**, 409–424 (1999).
106. Quadir, S. et al. Short- and long-range cation disorder in $(Ag_xCu_{1-x})_2ZnSnSe_4$ kesterites. *Chem. Mater.* **34**, 7058–7068 (2022).
107. Morrow, J. D. et al. Understanding defects in amorphous silicon with million-atom simulations and machine learning. *Angew. Chem. Int. Ed.* **63**, e202403842 (2024).
108. Riebesell, J. et al. Matbench discovery—an evaluation framework for machine learning crystal stability prediction. Preprint at https://arxiv.org/html/2308.14920v2 (2023).
109. Shimizu, K. et al. Using neural network potentials to study defect formation and phonon properties of nitrogen vacancies with multiple charge states in GaN. *Phys. Rev. B* **106**, 054108 (2022).
110. Ko, T. W., Finkler, J. A., Goedecker, S. & Behler, J. General-purpose machine learning potentials capturing nonlocal charge transfer. *Acc. Chem. Res.* **54**, 808–817 (2021).
111. Kavanagh, S. R., Scanlon, D. O., Walsh, A. & Freysoldt, C. Impact of metastable defect structures on carrier recombination in solar cells. *Faraday Discuss.* **239**, 339–356 (2022).
112. Mosquera-Lois, I., Kavanagh, S. R., Klarbring, J., Tolborg, K. & Walsh, A. Imperfections are not 0 K: free energy of point defects in crystals. *Chem. Soc. Rev.* **52**, 5812–5826 (2023).
113. Pols, M., Brouwers, V., Calero, S. & Tao, S. How fast do defects migrate in halide perovskites: insights from on-the-fly machine-learned force fields. *Chem. Commun.* **59**, 4660–4663 (2023).
114. Freysoldt, C. et al. First-principles calculations for point defects in solids. *Rev. Mod. Phys.* **86**, 253–305 (2014).
115. Lany, S. & Zunger, A. Assessment of correction methods for the band-gap problem and for finite-size effects in supercell defect

calculations: Case studies for ZnO and GaAs. *Phys. Rev. B* **78**, 235104 (2008).
116. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened coulomb potential. *J. Chem. Phys.* **118**, 8207–8215 (2003).
117. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
118. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).
119. Kresse, G. & Hafner, J. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B* **49**, 14251–14269 (1994).
120. Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Comput. Mater. Sci.* **111**, 218–230 (2016).
121. Uhrin, M., Huber, S. P., Yu, J., Marzari, N. & Pizzi, G. Workflows in AiiDA: engineering a high-throughput, event-based engine for robust and modular computational workflows. *Comput. Mater. Sci.* **187**, 110086 (2021).
122. Huber, S. P. et al. AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Sci. Data* **7**, 300 (2020).
123. Ong, S. P. et al. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
124. Jain, A. et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
125. Ong, S. P. et al. The materials application programming interface (API): a simple, flexible and efficient API for materials data based on REpresentational state transfer (REST) principles. *Comput. Mater. Sci.* **97**, 209–215 (2015).
126. Shen, J.-X. & Varley, J. pymatgen-analysis-defects: A python package for analyzing point defects in crystalline materials. *J. Open Source Softw.* **9**, 5941 (2024).
127. Shen, J.-X., Voss, L. F. & Varley, J. B. Simulating charged defects at database scale. *J. Appl. Phys.* **135**, 145102 (2024).
128. Larsen, A. H. et al. The atomic simulation environment-a python library for working with atoms. *J. Condens. Matter Phys.* **29**, 273002 (2017).
129. Kavanagh, S. R. et al. doped: Python toolkit for robust and repeatable charged defect supercell calculations. *J. Open Source Softw.* **9**, 6433 (2024).
130. Bitzek, E., Koskinen, P., Gähler, F., Moseler, M. & Gumbsch, P. Structural relaxation made simple. *Phys. Rev. Lett.* **97**, 170201 (2006).
131. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
132. Hinton, G. E. & Roweis, S. Stochastic neighbor embedding. *Adv. Neural Inf. Process Syst.* **15**, 857–864 (2002).
133. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).

Acknowledgements

The authors thank David O. Scanlon for discussions on defect symmetry breaking. I.M.L. acknowledges Imperial College London for funding a President's PhD scholarship. S.R.K. acknowledges the EPSRC Centre for Doctoral Training in the Advanced Characterisation of Materials (CDT-ACM) (EP/S023259/1) for funding a PhD studentship. A.M.G. is supported by EPSRC Fellowship EP/T033231/1. A.W. is supported by EPSRC project EP/X037754/1. We are grateful to the UK Materials and Molecular Modelling Hub for computational resources, which are partially funded by EPSRC (EP/P020194/1 and EP/T022213/1). This work used the ARCHER2 UK National Supercomputing Service (https://www.archer2.ac.uk) via our membership of the UK's HEC Materials Chemistry Consortium, which is funded by EPSRC (EP/L000202). We acknowledge the Imperial College London's High Performance Computing services for computational resources.

Author contributions

Conceptualisation & Project Administration: All authors. Investigation and methodology: I.M.-L. Supervision: S.R.K., A.M.G., A.W. Writing—original draft: I.M.-L. Writing—review & editing: All authors. Resources and funding acquisition: A.M.G., A.W. These author contributions are defined according to the CRediT contributor roles taxonomy.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-024-01303-9.

Correspondence and requests for materials should be addressed to Aron Walsh.

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

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

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

© The Author(s) 2024

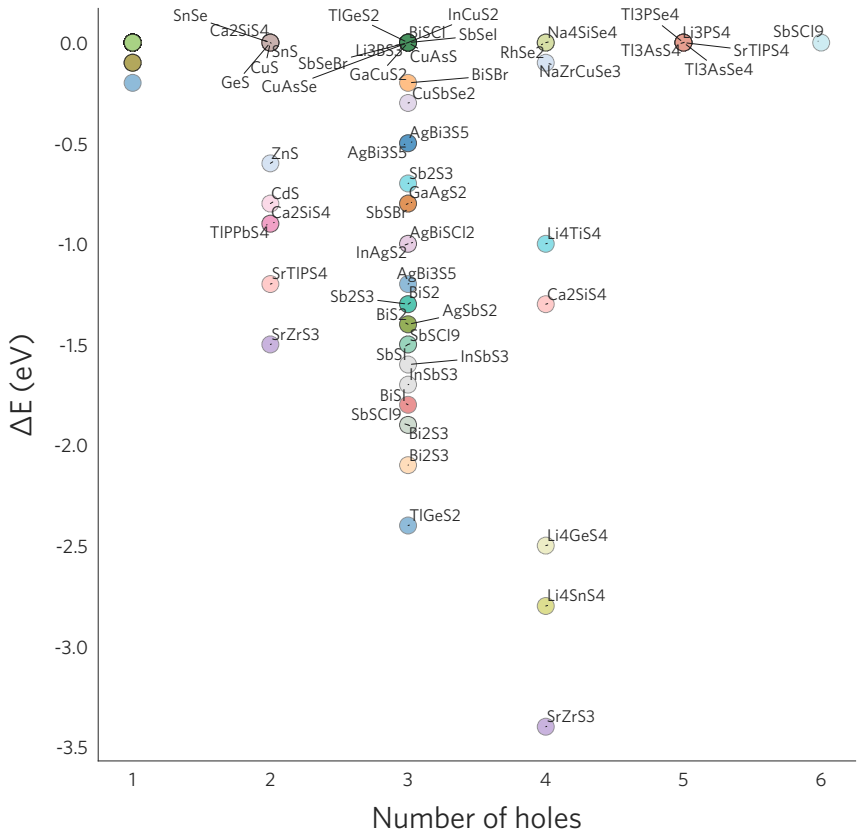
Supplementary Information for 'Machine-learning structural reconstructions for accelerated point defect calculations'

1. Supplementary Notes

A. Dataset analysis

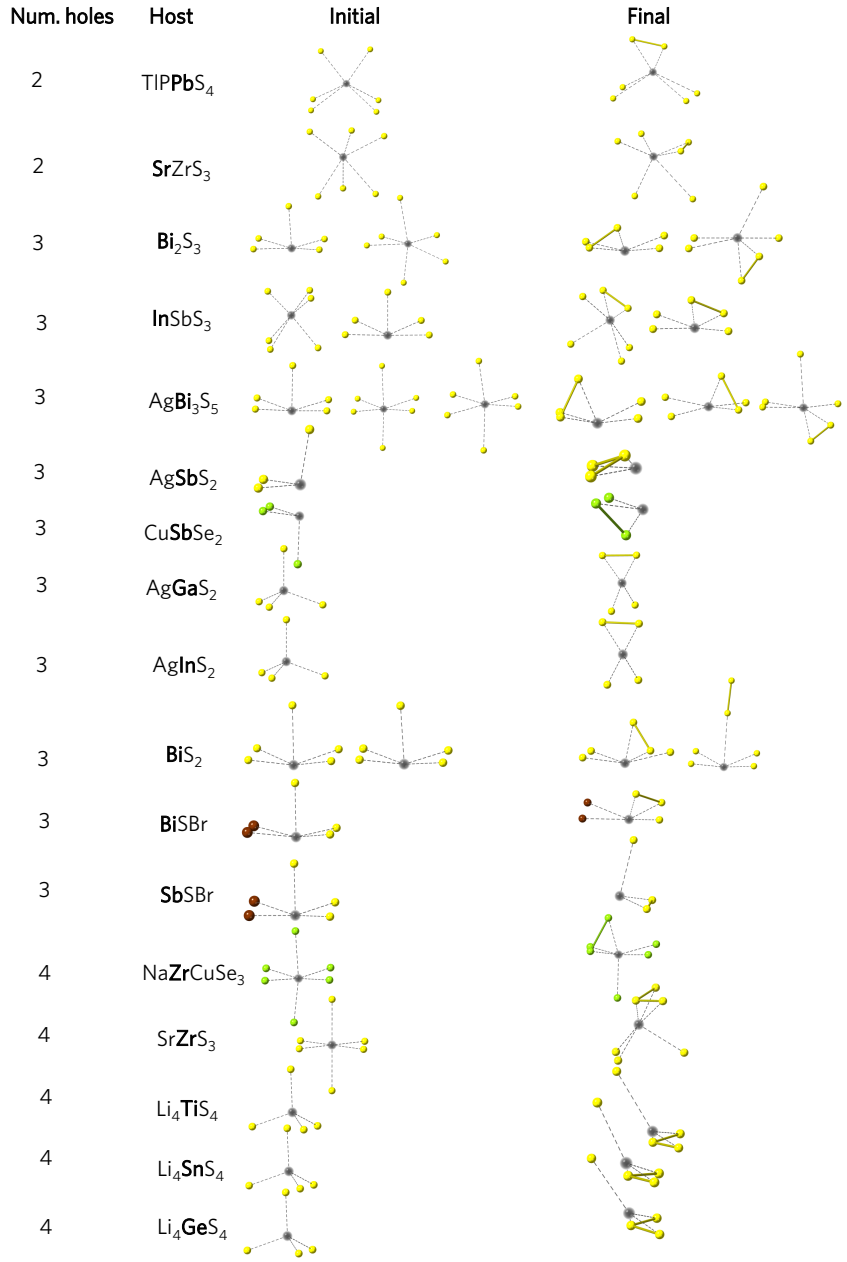
Supplementary Table 1: Percentage of defects that lead to energy-lowering reconstructions (compared to a typical unperturbed relaxation) higher than the specified energy threshold. The column 'Reconstructions found with rattling' indicates the percentage of reconstructions identified if we only apply randomised distortions to the unperturbed, ideal defect structure (i.e. no bond distortions), demonstrating the need for proper structure searching.

Threshold (eV)	Significant reconstructions (%)	Reconstructions found with rattling (%)
-0.05	40.3	3.7
-0.10	33.6	0
-0.50	29.9	0
-1.00	20.9	0
-2.00	6.7	0

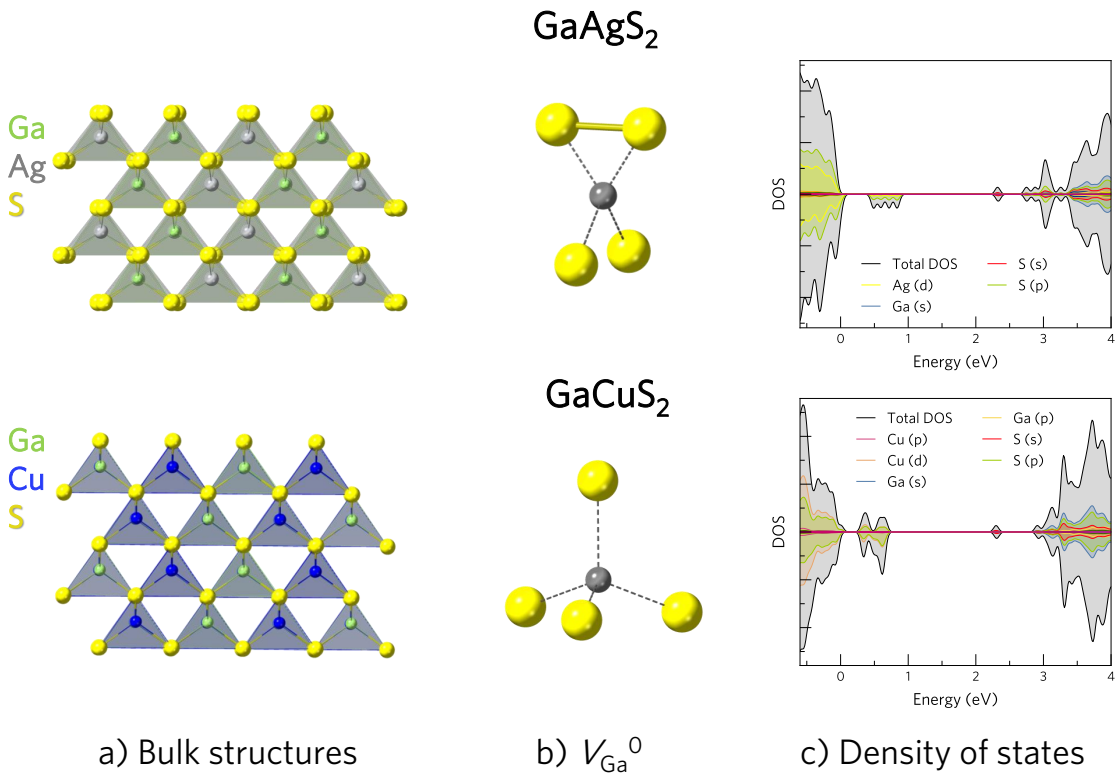


Supplementary Figure 1: Energy difference between the ground state structure identified with ShakeNBreak and the final structure obtained by relaxing the ideal, high-symmetry defect structure for cation vacancies across each chalcogenide. Note that all defects with more than 3 missing electrons form dimer reconstructions, but some appear with $\Delta E = 0$ eV since the ground state structure is already identified from the high-symmetry relaxation. We note that there is a high overlap of data points with 1 missing electron at $\Delta E = 0$ eV, since most defects with 1 missing electron do not lead to favourable reconstructions. These points are not labelled to avoid overlap of labels and improve readability.

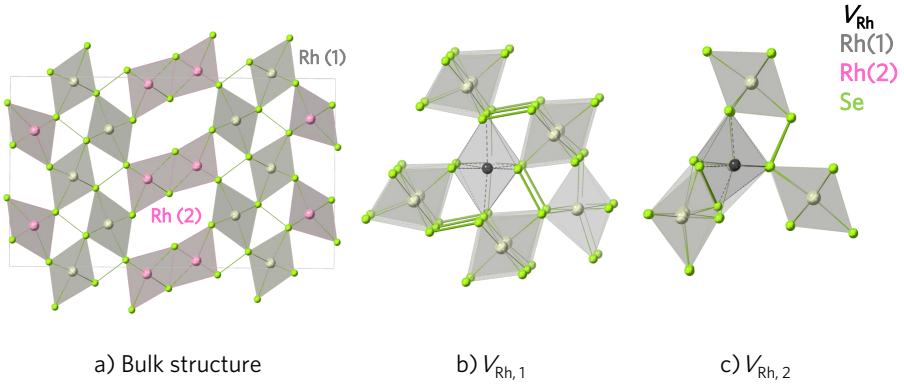
1. *Defect reconstructions*



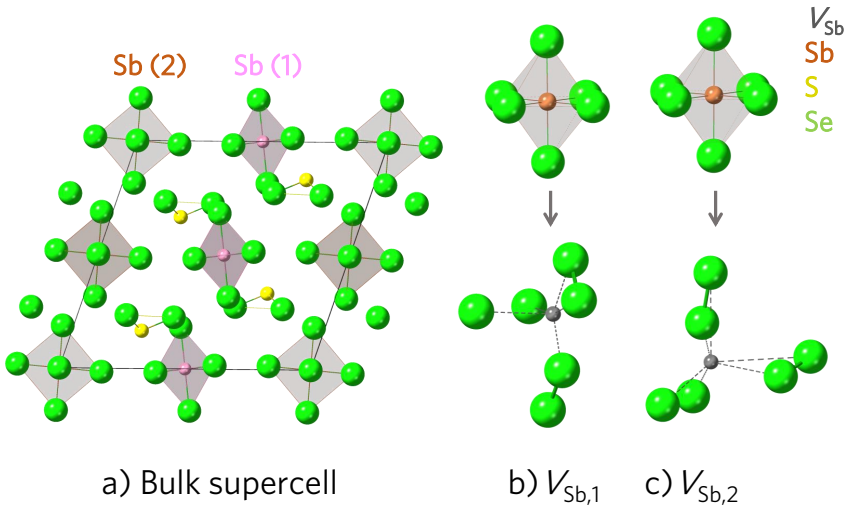
Supplementary Figure 2: Illustration of some of the reconstructions identified, showing the similar motifs undergone by different hosts. The initial (high-symmetry) and the ground state structures are shown. The defect element is shown in bold in the host composition.



Supplementary Figure 3: Reconstructions undergone by $V_{Ga}{}^{0}$ in $GaAgS_2$ and $GaCuS_2$. a) Bulk structures for $GaAgS_2$ (top) and $GaCuS_2$ (bottom). b) $V_{Ga}{}^{0}$, showing the dimerisation undergone in $GaAgS_2$ while in $GaCuS_2$ $V_{Ga}{}^{0}$ maintains the ideal tetrahedral coordination. Similar behaviour is observed for $V_{In}{}^{0}$ in $InAg/CuS_2$. c) Density of states for $V_{Ga}{}^{0}$, showing the localised defect states within the gap.



Supplementary Figure 4: Reconstructions undergone by the symmetry inequivalent Rh vacancies in $RhSe_2$. a) Bulk structure showing the two symmetry inequivalent Rh sites, in grey (Rh(1)) and pink (Rh(2)). b) $V_{\mathrm{Rh},1}$ does not form a Se dimer since the holes can localise in the Se-Se bonds already neighbouring this site. c) $V_{\mathrm{Rh},2}$ forms an additional Se-Se bond as it lacks enough neighbouring Se-Se bonds to localise the four holes.



Supplementary Figure 5: Reconstructions undergone by the symmetry inequivalent Sb vacancies in $SbSCl_9$. a) Bulk structure showing the two symmetry inequivalent Sb sites, in pink (Sb(1)) and brown (Sb(2)). b, c) $V_{\mathrm{Sb},1}$ and $V_{\mathrm{Sb},2}$ forming two Cl-Cl dimers.

B. Model training

In this section, we compare different training strategies and parameters. We note that errors are higher than in the final model since here most of the parameters were set to default values except for the parameter being benchmarked. Further, in these experiments, the validation set only contained unseen compositions (and not unseen configurations from compositions included in the training set).

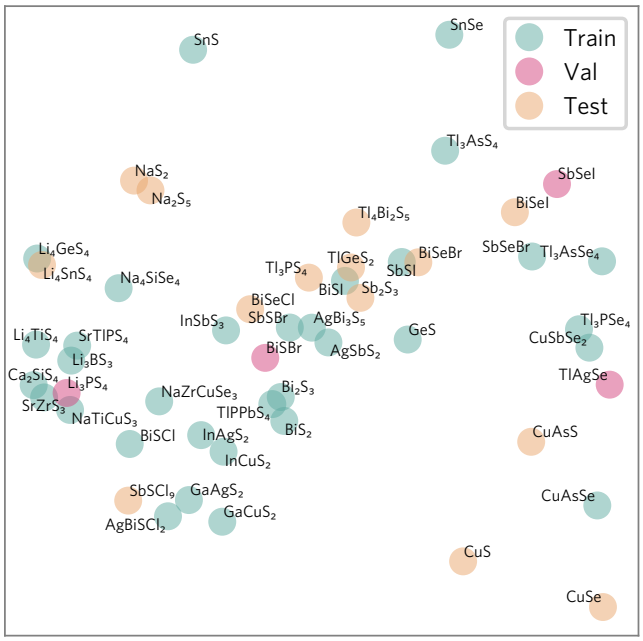
1. Splits

The dataset was first split by composition into training, validation, and test sets in a balanced way (i.e. balanced distribution of the constituent elements and similar host systems (e.g. Li_4SnS_4, Li_4GeS_4, Li_4TiS_4)), as shown in Fig. 6.

2. Reference energies

Before fitting most MLFFs, it is common to subtract the reference elemental energies from the total energies to improve training stability[80]. In M3GNet[80], the reference elemental energies can be calculated from the training or other user-specified data using linear regression[80]. Since the original M3GNet model was trained on GGA data using GGA values for the elemental reference energies, it is important to update the reference energies with HSE values to correct the systematic differences between the energies calculated with the two functionals. To compare the effect of the reference elemental energies, we considered three different approaches to calculate them: i) energies of the isolated atoms, ii) linear regression[a] to the bulk energies, and iii) linear regression to the energies of all training defect structures. As expected, the last case decreases performance since the regressed reference energies include defect formation contributions. On the other hand, when the reference elemental energies are calculated from the bulk systems, these constitute a robust reference that simplifies learning the energies of different defect structures.

[a] The linear regression was performed using the `AtomRef` class from Ref. 80.



Supplementary Figure 6: Plot of the first two principal components of the feature space for the pristine structures, showing the division into training, validation, and test sets. We note that the validation set was then augmented with some configurations from the compositions in the training set. The structures were encoded with the 128-element vector outputs from the M3GNet model trained on the formation energies of bulk materials in the Materials Project database, as done in Ref. 79.

Supplementary Table 2: Comparison of model performance depending on the elemental reference energies used. The same training and validation data and training parameters are used for all cases and the performance is measured with the mean absolute errors of the energies, forces, and stresses.

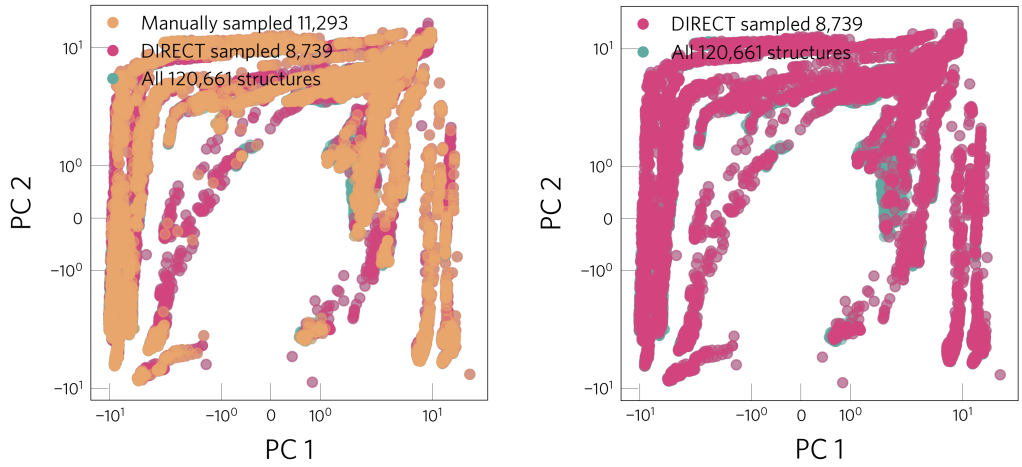
Method	$Loss_{val}$	E_{val} (meV/atom)	F_{val} (meV/Å)	S_{val} (GPa)	E_{train} (meV/atom)	F_{train} (meV/Å)	S_{train} (GPa)
Bulk	0.22	41.8	153.8	0.22	50.50	118.0	0.19
Isolated atoms	0.31	72.4	191.6	0.43	134.5	154.3	0.30
Defect	0.34	130.3	179.3	0.34	109.6	158.2	0.28

3. Sampling methods

To investigate how to best sample configurations from the relaxation trajectories to generate the training set we compared two approaches: i) a manual method where we sample 10 evenly spaced frames from each relaxation (MS) and ii) the Dimensionality-Reduced Encoded Clusters approach (DIRECT)[79], which aims to select a robust training set from a complex configurational space (Fig. 7). This comparison was performed without including configurations of the bulk (pristine) host structures. After generating the different training sets, four M3GNet models[80] with default parameters were trained and evaluated. As shown in Table 3, when using training sets of similar sizes, the manual model performs better. This results from the DIRECT approach mostly sampling highly distorted configurations from the initial ionic steps (Fig. 9), hindering the learning of the low-energy region of the potential energy surface (PES).

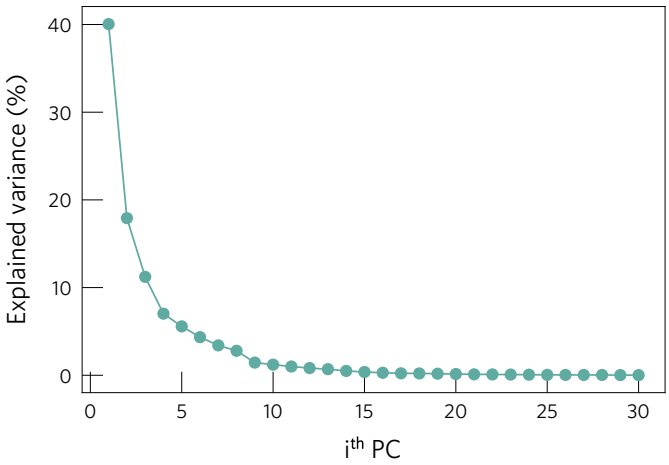
Supplementary Table 3: Comparison of the mean absolute errors in the validation and training sets obtained with the different methods of sampling the training data. The same validation data and training parameters are used for all cases. The column 'Params.' lists the sampling parameters used with the BIRCH clustering method, as detailed in Ref. 79.

Method	Size	Params. (t, n, k)	$Loss_{val}$	E_{val} (meV/atom)	F_{val} (meV/Å)	S_{val} (GPa)	E_{train} (meV/atom)	F_{train} (meV/Å)	S_{train} (GPa)
Direct	21149	0.02, 4000, 20	0.20	28.6	148.6	0.19	38.1	147.6	0.19
Manual	11335	-	0.22	51.4	146.7	0.26	62.0	123.3	0.23
Direct	8705	0.1, 1000, 20	0.25	47.0	174.7	0.33	42.2	177.3	0.23
Direct	13144	0.03, 2000, 20	0.28	93.4	165.3	0.26	54.9	192.3	0.26

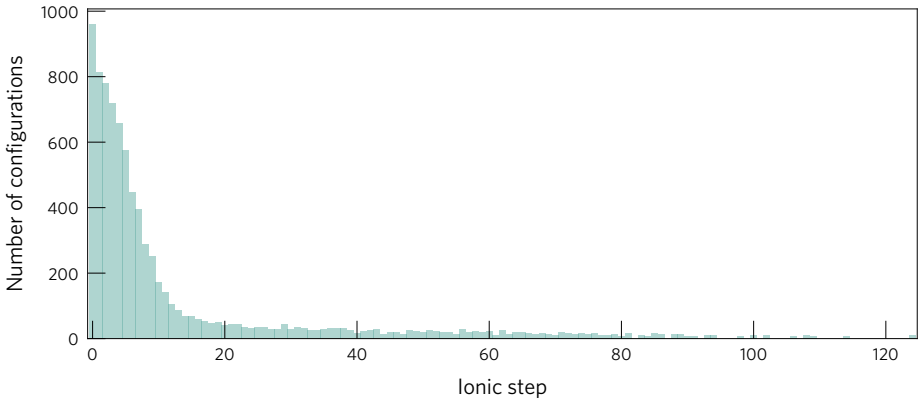


(a) DIRECT and manual coverage (b) DIRECT only (for clarity)

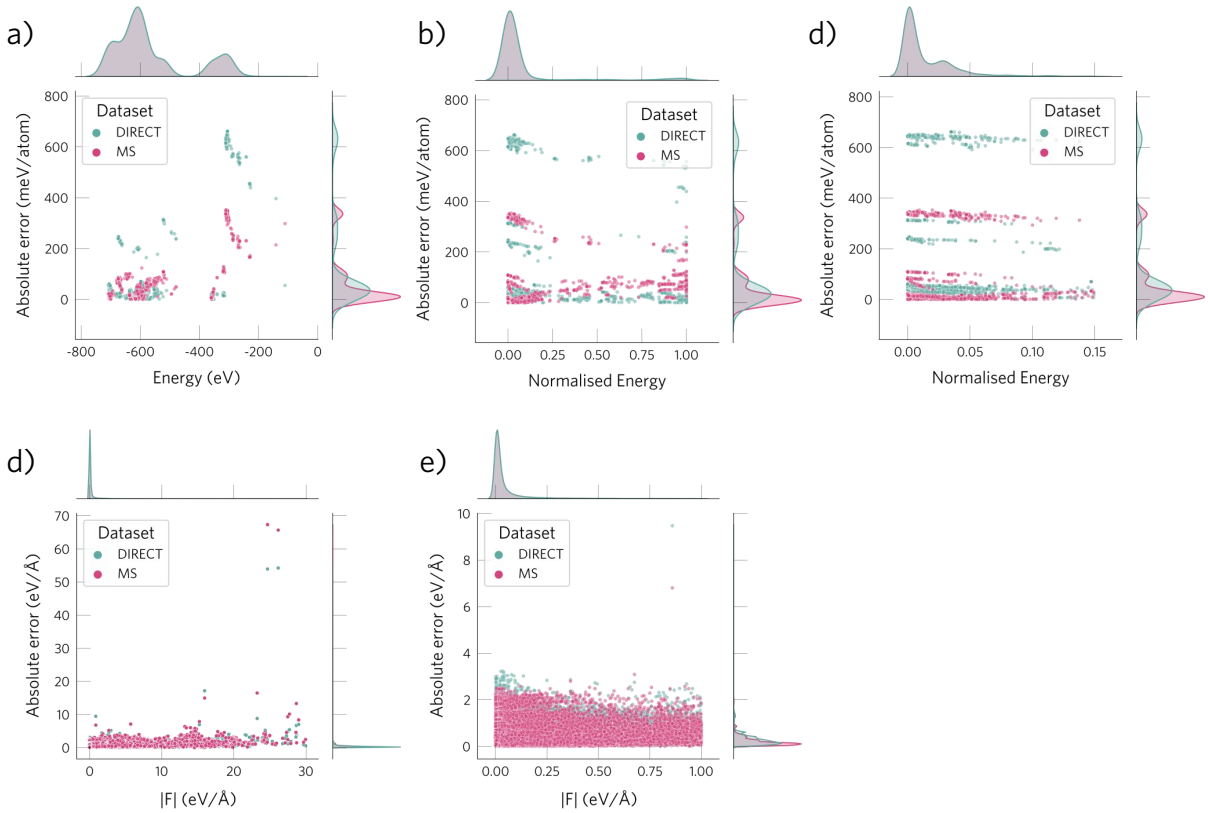
Supplementary Figure 7: Sampling coverage of manual (orange) and DIRECT sampling (pink) of the full dataset (green), illustrating that the DIRECT approach covers better the configurational landscape. For the DIRECT method, we considered the parameters t=0.02, n=4000, k=20, but similar behaviour is observed for the other cases of Table 3. Note that the dataset is only plotted along two dimensions (see Fig. 8). The structures were encoded with the 128-element vector outputs from the M3GNet model trained on the formation energies of bulk materials in the Materials Project database[79]. For further details of the method used to featurise structures, dimensionality reduction, clustering, and stratified sampling see the original publication of Qi *et al*[79].



Supplementary Figure 8: Explained variance versus the number of PCA dimensions. PCA was used to reduce the dimensions of the feature vector used to describe the defect configurations (see Ref. 79 for details).



Supplementary Figure 9: Distribution of DIRECT (t=0.02, n=4000, k=20) sampled configurations, showing that structures are mainly selected from the initial ionic steps, which correspond to highly distorted structures.



Supplementary Figure 10: Distribution of the energy (a, b, c) and force (d, e) absolute errors for the test set. The MS sampling results in lower errors for the low-energy structures while the DIRECT approach leads to higher errors for these structures. a) Distribution of absolute energy errors versus the absolute DFT energy for all test compositions. b) Same as (a) but with the DFT energies normalised for each defect, so that energies of different defects/compositions are comparable. c) Same as (b) but removing very high energy configurations to aid visualisation. d) Distribution of absolute force errors versus the magnitude of the DFT forces. Same as (c) but removing configurations with very high forces. We note that the configurations with very high errors correspond to the most challenging compositions (more different from the training set, like the quasi-one-dimensional system $SbSCl_9$).

4. Model architecture

To compare different model architectures, we trained from scratch (i.e. initialising the weights with random values) three models: M3GNet[80], MACE[84] and CHGNet[81]. Again, for simplicity, this comparison was performed with default parameters for each model. The same training, validation, and test sets were used to compare the different models. The worse performance of CHGNet compared to the other models seems to arise from two reasons: i) the smaller value of the 3-body cutoff that is used by default (3 Å compared to 4 Å in M3GNet) and ii) its default readout or pooling layer, which involves an average function (with the attention layer performing better). Regarding M3GNet and MACE, MACE seems to perform better on the test set. However, due to the lack of a pre-trained MACE universal model at the time when this comparison was performed, we decided to use M3GNet. However, with the now available MACE universal model, we expect the MACE architecture to perform better.

Supplementary Table 4: Comparison of validation and test mean absolute errors for different models (M3GNet, CHGNet, MACE). For the test set, the Spearman coefficient for the energies (ρ) is also shown. The default parameters of each model were used and the models were trained from scratch (without fine-tuning/pretraining). In all cases, the training data included energies, forces and stresses.

Model	E_{val} (meV/atom)	F_{val} (meV/Å)	E_{test} (meV/atom)	F_{test} (meV/Å)	ρ_{test}
M3GNet	45.8	143.1	129.2	263.0	0.44
MACE[a]	47.8	173.9	153.7	187.8	0.66
CHGNet	192.3	351.4	376.0	333.8	0.37

[a] For MACE, the following parameters were used: hidden irreducible representations: 32x0e, radial cutoff: 5 Å; weight for energies, forces and stresses: 1, 1, 0.1; learning rate scheduler: exponential; batch size: 20; maximum number of epochs: 250.

5. Training parameters for M3GNet model

To determine the best training parameters for the M3GNet model, we performed several benchmarks by comparing training and validation errors. From these comparisons, we determined the best learning rate, radial and 3-body cutoffs and readout function. We note that these comparisons were performed on the training and validation errors, rather than test ones, as they were intended as quick benchmarks to determine appropriate training parameters. Further, note that the errors are higher than for the final model since the default values were used for the parameters not being tested in each experiment.

a Learning rate

Supplementary Table 5: Comparison of validation mean absolute errors for different learning rates and learning rate schedulers. 'Constant' denotes using a constant learning rate. When a scheduler is used, the initial learning rate is given in parentheses. Rows are ordered by ascending validation loss. The same training data (manual sampling) and training parameters were used for all experiments.

Method	Loss_{val}	E_{val} (meV/atom)	F_{val} (meV/Å)	S_{val} (GPa)	E_{train} (meV/atom)	F_{train} (meV/Å)	S_{train} (GPa)
Exp. scheduler ($5 \cdot 10^{-4}$)	0.19	35.7	128.0	0.22	30.9	93.0	0.16
Cos. scheduler (10^{-3})	0.20	41.8	135.4	0.26	32.2	106.8	0.17
Constant ($5 \cdot 10^{-4}$)	0.21	45.0	142.6	0.19	34.1	96.2	0.16
Constant (10^{-3})	0.21	45.8	143.1	0.19	36.7	97.9	0.17
Exp. scheduler (10^{-3})	0.21	25.8	152.9	0.36	21.8	85.70	0.12
Constant (10^{-4})	0.24	58.8	152.1	0.29	56.7	131.4	0.22
Time. scheduler (10^{-3})	0.25	64.3	152.0	0.31	35.2	101.4	0.16
Constant (10^{-5})	0.29	70.9	191.2	0.32	84.8	174.1	0.41

b Structure featurisation

We performed benchmarks to identify the optimal values for the radial and 3-body cutoff and the optimal function for the pooling layer. As shown in Tables 6 to 8, we found a 3-body cutoff of 4 Å, a radial cutoff of 5 Å, and the weighted atom layer from the M3GNet model to work best.

Supplementary Table 6: Comparison of cutoff values for the 3-body radius in the M3GNet model. Mean absolute errors for energies, forces, and stresses for the training and validation sets. Rows are ordered by ascending validation loss.

cutoff (Å)	Loss_{val}	E_{val} (meV/atom)	F_{val} (meV/Å)	S_{val} (GPa)	E_{train} (meV/atom)	F_{train} (meV/Å)	S_{train} (GPa)
4	0.19	35.7	128.0	0.22	30.9	93.0	0.16
3	0.26	59.2	164.9	0.32	50.6	123.5	0.20

Supplementary Table 7: Performance comparison for different radial cutoffs in the M3GNet model. Mean absolute errors for energies, forces, and stresses for the training and validation sets. Rows are ordered by ascending validation loss.

cutoff (Å)	Loss_{val}	E_{val} (meV/atom)	F_{val} (meV/Å)	S_{val} (GPa)	E_{train} (meV/atom)	F_{train} (meV/Å)	S_{train} (GPa)
5	0.19	35.7	128.0	0.22	30.9	93.0	0.16
4.5	0.22	40.1	142.0	0.34	35.1	97.4	0.16
5.5	0.23	57.1	140.7	0.31	46.8	114.4	0.20
6	0.23	73.9	133.0	0.22	45.9	103.4	0.18

Supplementary Table 8: Performance comparison for different readout functions in the M3GNet model. Mean absolute errors for energies, forces, and stresses for the training and validation sets. Rows are ordered by ascending validation loss.

Readout	Loss_{val}	E_{val} (meV/atom)	F_{val} (meV/Å)	S_{val} (GPa)	E_{train} (meV/atom)	F_{train} (meV/Å)	S_{train} (GPa)
Weighted atom	0.19	35.7	128.0	0.22	30.9	93.0	0.16
Reduce readout	0.23	54.6	143.1	0.30	65.0	135.1	0.24
Set2Set	0.26	57.7	163.9	0.39	44.3	130.9	0.2

c Fine-tuning

To assess the effect of fine-tuning the model from a model trained on the bulk relaxations of the Materials Project database, we compared the performance of four different models: training from scratch (i.e. initialising the weights with random values), or fine-tuning the bulk model (either training all layers or only the last final layers). When comparing their performance on the validation set, we found that fine-tuning the bulk model and training all layers resulted in the best performance, with the main benefit observed for the forces.

Supplementary Table 9: Performance comparison between different training strategies: training from scratch (initialising the weights with random values), or fine-tuning the model previously trained on the Materials Project dataset of bulk structures[80] (with either re-training all layers (all), only the last two layers (2) or only the last layer (1)). The performance is measured with the mean absolute errors in the validation and training sets.

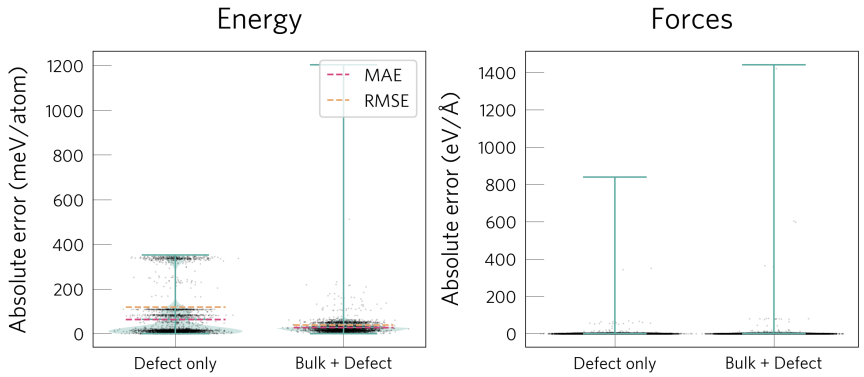
Training	Loss_{val}	E_{val} (meV/atom)	F_{val} (meV/Å)	S_{val} (GPa)	E_{train} (meV/atom)	F_{train} (meV/Å)	S_{train} (GPa)
Fine-tuning (all)	0.157	42.6	101.1	0.1	18.7	68.1	0.1
Fine-tuning (1)	0.176	37.6	118.8	0.2	18.3	76.0	0.1
Fine-tuning (2)	0.183	36.5	104.9	0.4	15.2	55.4	0.1
From scratch	0.187	46.4	121.5	0.2	27.7	86.1	0.1

d Adding bulk data to enhance defect dataset

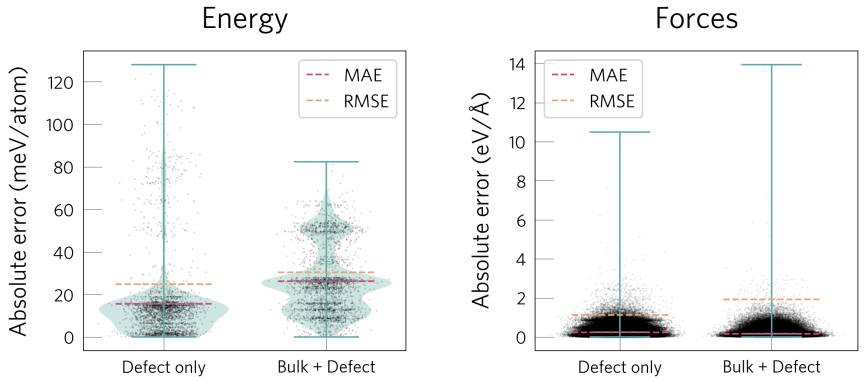
To test whether adding data for pristine systems would improve model performance, we trained two models: one on just the defect dataset (using MS sampling) and another model using the same defect dataset but combined with a small fraction of pristine configurations for the same compositions (5 evenly spaced frames from the relaxation of each pristine structure). As shown in Table 10, adding bulk data reduces the mean absolute errors for the energies and the forces of the test configurations. By analysing the error distributions in Fig. 11, we see that the main benefit of including bulk data is to reduce the errors for systems that are difficult due to their higher structural difference from the training set (i.e. including bulk data results in lower MAE and RMSE when considering all test systems (Fig. 11.a) but in a higher MAE and RMSE when not considering the harder compositions (Fig. 11.b)).

Supplementary Table 10: Performance comparison when adding bulk (pristine) configurations to the training dataset. Mean absolute errors (MAE) and root mean square errors (RMSE) for the (defect) test set when only training on defect data and when training on both defect and bulk data.

Dataset	MAE_E (meV/atom)	$RMSE_E$ (meV/atom)	ρ	MAE_F (meV/Å)	$RMSE_F$ (meV/Å)	MAE_S (GPa)	$RMSE_S$ (GPa)
Defect + Bulk	27.3	39.4	0.70	86.8	1943.5	0.19	0.54
Defect only	63.4	119.5	0.63	135.0	1139.6	0.34	0.70

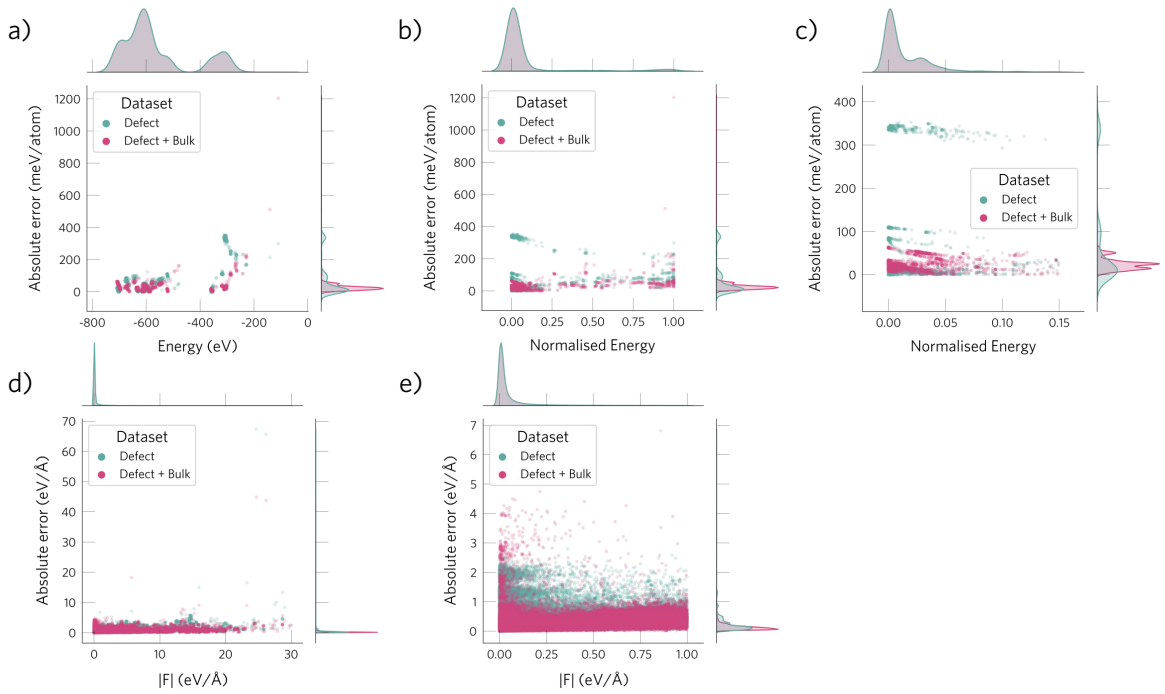


(a) All test systems

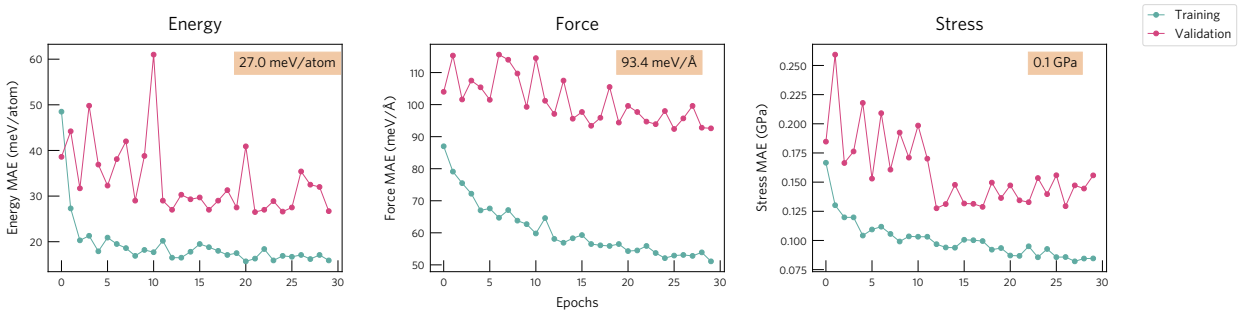


(b) Removing difficult compositions from test set

Supplementary Figure 11: Violin plots[?] and error metrics for energies and forces on the test set. a) When considering all test systems. b) When filtering out the systems that are hard to learn due to high structural differences to the training set ($SbSCl_9$, NaS_2 and CuS).



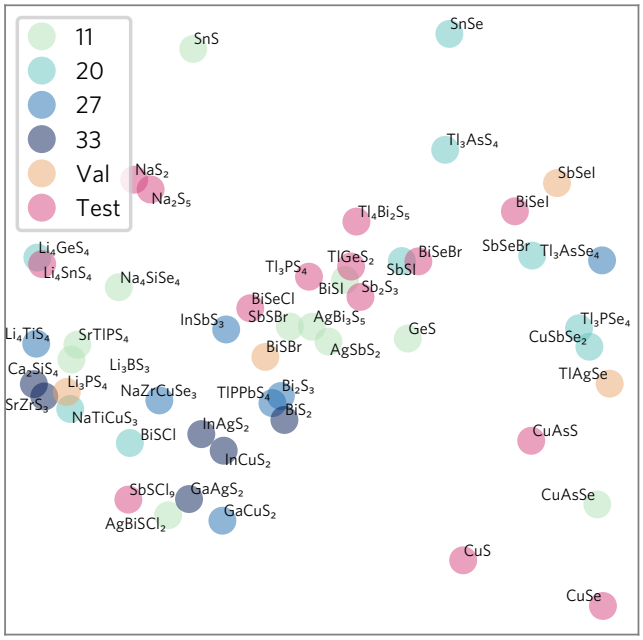
Supplementary Figure 12: Distribution of absolute errors for energies (a, b, c) and forces (c, d) when training on only defect data (green) and both defect and bulk data (pink). a) Distribution of absolute energy errors versus the absolute DFT energy for all test compositions. b) Same as (a) but with the DFT energies normalised for each defect, so that energies of different defects/compositions are comparable. c) Same as (b) but removing very high energy configurations, illustrating how adding bulk data reduces the errors of many low-energy structures. d) Distribution of absolute force errors versus the magnitude of the DFT forces. Same as (c) but removing configurations with very high forces, illustrating how adding bulk data reduces the errors for low force configurations.



Supplementary Figure 13: Evolution of training and validation errors with epoch number. The validation errors obtained for the epoch with the lowest validation loss are shown at the top right of each subplot.

6. *Learning curve*

To investigate the dependence of the performance on training set size, we generated four training sets with increasing number of compositions. As shown in Fig. 14, each set was generated by selecting diverse compositions (e.g. avoiding sets with compositions very close in the 2D feature map). Each of these sets was used to retrain the universal M3GNet model, which was then applied to the defects in the test set. For comparison, the universal M3GNet model (without retraining) was also applied. As shown in Table 11, the main benefit of increasing the training set size is to reduce the mean absolute error of the forces. However, we note that these results depend on the specific splits (e.g. which compositions are included in the different sets) and likely require a larger test set to properly evaluate the dependence on training set size.



Supplementary Figure 14: Distribution of the compositions included in the different training sets. The legend denotes the number of compositions included in each set. Each set is generated by adding the coloured compositions to the previous set (e.g. the set with 20 compositions includes the compositions shown in light green (included in the set-11) plus the ones shown in light blue).

Supplementary Table 11: Comparison of mean absolute errors on energies, forces and stresses for the four models trained on sets with increasing number of compositions. For comparison, the performance of the universal M3GNet model (without retraining to defect data) is also included, as well as the performance of the universal M3GNet model with updated reference elemental energies to their HSE06 values, which significantly reduces the errors.

Num. compositions	E (meV/atom)	F (meV/Å)	S (GPa)
MP	735.5	107.7	0.52
MP(ref)	249.2	107.7	0.52
11	23.2	54.0	0.16
20	26.7	54.5	0.13
27	22.7	59.6	0.19
33	27.5	50.6	0.15

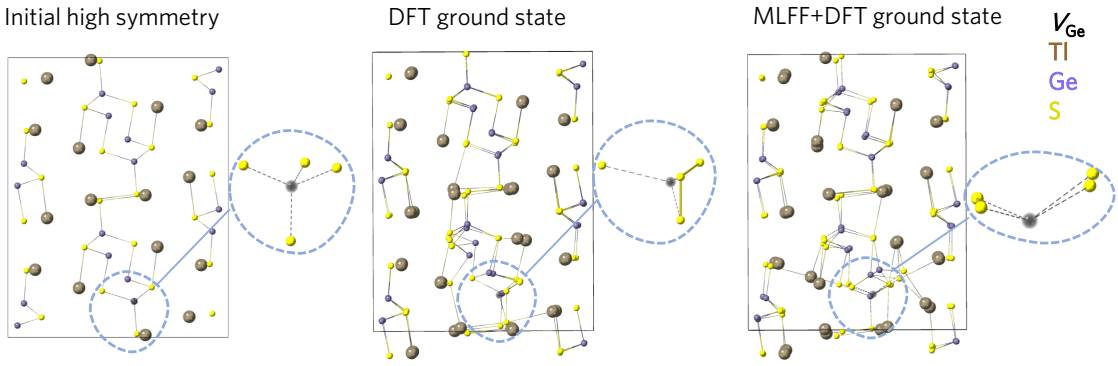
C. Model performance

1. Metrics

Supplementary Table 12: Performance of the fine-tuned surrogate model on the test set. Mean absolute errors (MAE, in eV and eV/atom) when predicting the relative energies of the low-energy configurations of a defect (i.e. structures that are less than 5 eV above the defect ground state configuration).

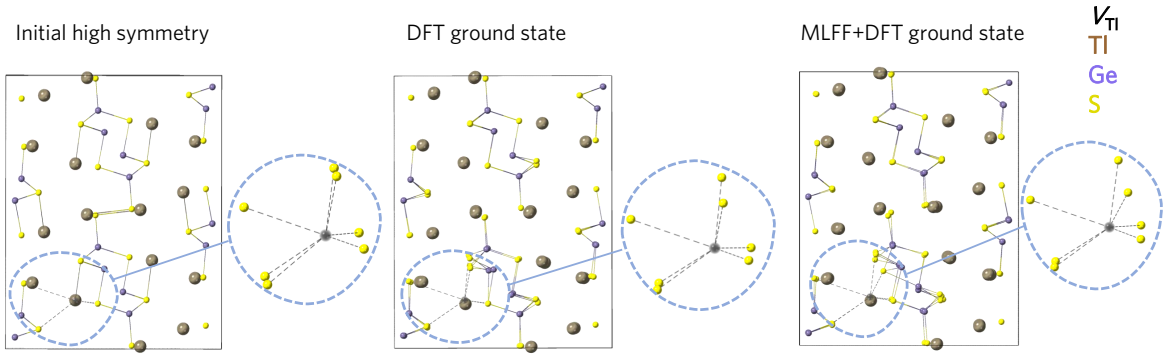
	MAE (eV)	MAE (meV/atom)
Li_4SnS_4	1.6	11.2
BiSeBr	0.7	1.4
$TlGeS_2$	0.6	0.7
Tl_3PS_4	0.4	1.6
$SbSCl_9$	0.2	9.6
$Tl_4Bi_2S_5$	0.2	5.0
BiSeCl	0.2	1.2
CuS	0.2	1.2
Na_2S_5	0.2	2.9
CuAsS	0.1	0.9
NaS_2	0.1	2.8
CuSe	0.1	1.9

2. New ground states: $TlGeS_2$

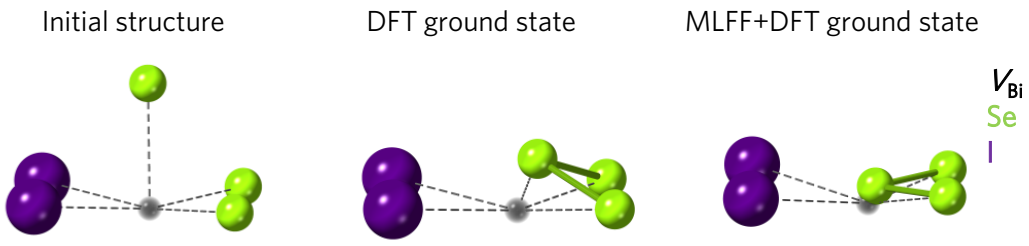


Supplementary Figure 15: $V_{Ge,9}$ in $TlGeS_2$. Ground state structures identified with the DFT search (S trimer) and with the MLFF+DFT approach (*one* S dimer), with the second being 0.5 eV lower in energy. The structure with only one S-S bond seems to be more favourable due to avoiding strain caused by the rearrangement of the atoms. For clarity, the initial high symmetry defect structure is shown on the left.

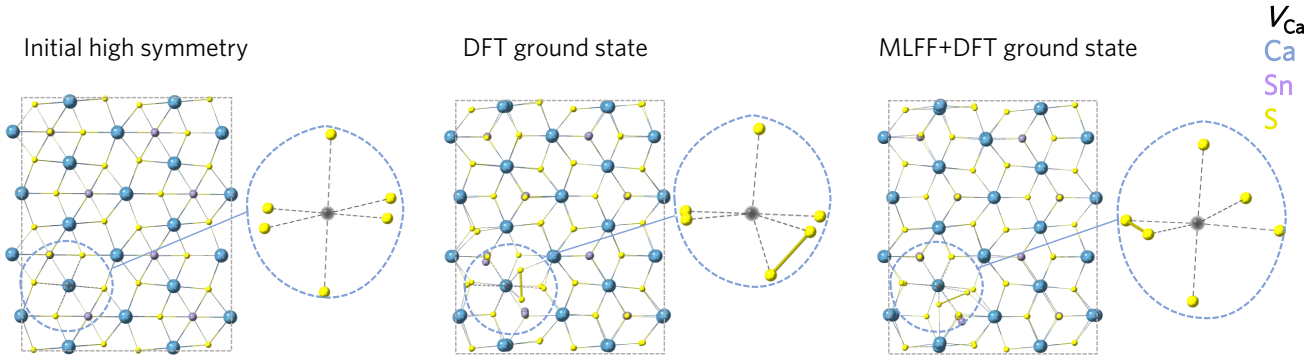
3. Failed systems



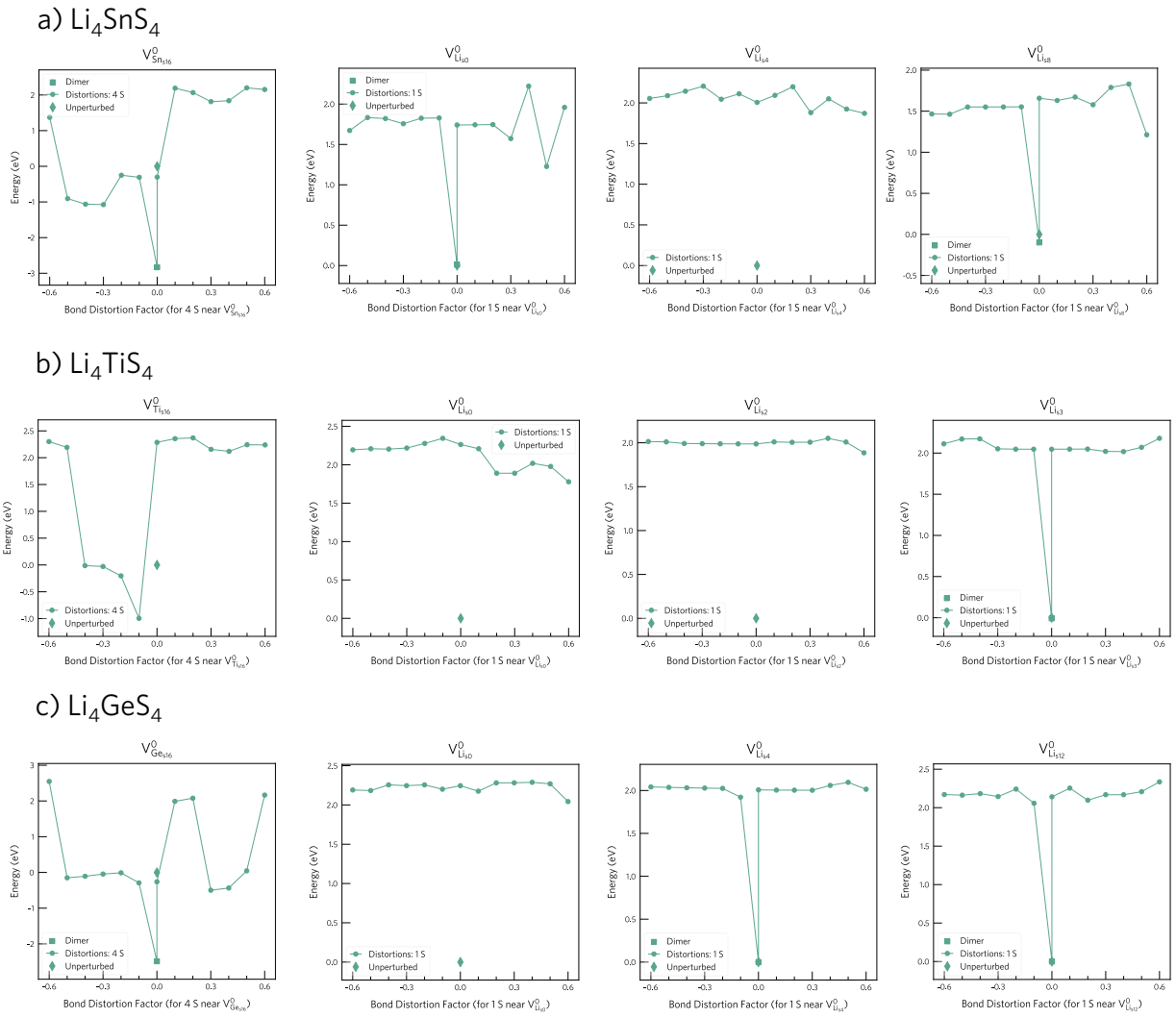
Supplementary Figure 16: High symmetry (initial) and ground state structures of $V_{Tl,0}$ in $TlGeS_2$ identified with the DFT-only search and the MLFF+DFT approach. While the configuration identified with the DFT search is 0.1 eV lower in energy than the MLFF+DFT structure, we note that the reconstruction motifs are very similar.



Supplementary Figure 17: High symmetry (initial) and ground state structures of V_{Bi} in BiSeI identified with the DFT-only search and the MLFF+DFT approach. While the configuration identified with the DFT search is 0.2 eV lower in energy than the MLFF+DFT structure, we note that the reconstruction motifs are similar, involving a Se trimer in both cases.



Supplementary Figure 18: High symmetry (initial) and ground state structures of $V_{\mathrm{Ca},0}$ in Ca_2SnS_4 identified with the DFT-only search and the MLFF+DFT approach. While the configuration identified with the DFT search is 0.7 eV lower in energy than the MLFF+DFT structure, we note that the reconstruction motifs are similar, involving a S dimer in both cases.



Supplementary Figure 19: Energy vs distortion plots for all the cation vacancies in the test system Li_4SnS_4 (a) and the training compositions Li_4TiS_4 and Li_4GeS_4 (b, c). Note that most of the relaxations are trapped in high-energy minima, thus hindering learning the low-energy region of the PES for these systems. This trapping into high-energy basins can be avoided by reducing the magnitude of the rattle distortion that is applied with ShakeNBreak to the sampling structures.

4. Acceleration factors

Supplementary Table 13: Comparison of computing times for structure searching when using the full DFT approach and the MLFF+DFT approach. We compare the time required to run the structure searching relaxations when only using DFT (i.e. starting from the ShakeNBreak initial structures) and when pre-relaxing the structures with the MLFF. These timings are listed in the columns 'DFT time' (full DFT relaxation of ShakeNBreak structures), 'MLFF time' (DFT relaxation of the MLFF-relaxed ShakeNBreak structures) and 'Inference time' (MLFF relaxation of ShakeNBreak structures).

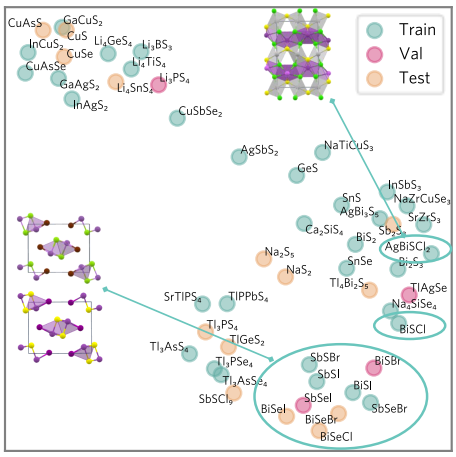
Host	Defect	Num atoms	Num electrons	DFT time (CPU h)	MLFF time (CPU h)	Inference time (GPU h)	Speedup[a]
BiSeBr	V_{Bi_0}	71	657	4736.7	3742.7	0.05	1.3
BiSeI	V_{Bi_0}	71	657	5778.6	3160.8	0.04	1.8
Li_4SnS_4	$V_{\mathrm{Sn}_1}6$	143	786	13976.7	18042.3	0.21	0.8
Li_4SnS_4	V_{Li_0}	143	797	11549.0	2576.1	0.18	4.5
Li_4SnS_4	V_{Li_4}	143	797	433083.1	6847.8	0.24	63.2
Li_4SnS_4	V_{Li_8}	143	797	41380.0	6205.6	0.15	6.7
CuSe	V_{Cu_0}	107	907	21196.8	969.0	0.06	21.9
CuSe	V_{Cu_4}	107	907	27734.9	1647.9	0.05	16.8
CuS	V_{Cu_0}	143	1213	30292.0	7821.3	0.03	3.9
CuS	V_{Cu_3}	143	1213	39892.4	3701.4	0.03	10.8

[a] The speedup factor is calculated as $\frac{t_{\mathrm{relax,DFT}}}{t_{\mathrm{relax,MLFF}}+t_{\mathrm{infer,MLFF}}}$, with a mean value of 13.2. The DFT relaxations of both the ShakeNBreak initial structures and the MLFF-relaxed structures are performed on 48 cores (node with two AMD EPYC 7742 processors). The MLFF relaxations of the ShakeNBreak initial structures are performed on a Quadro RTX6000 GPU.

5. Surrogate model for closely-related systems

To investigate whether targeting more similar systems would reduce the dataset size required for similar model performance, we developed a model for chalcohalide systems. We selected the chalcohalides from our dataset, resulting in 11 compositions (BiSCl, BiSBr, BiSI, BiSeCl, BiSeBr, BiSeI, SbSBr, SbSI, SbSeBr, SbSeI $AgBiSCl_2$) (Fig. 20). Three of

these (27%; BiSeCl, BiSeBr, BiSeI) were held out as the test set, and the remaining data was split into training and validation sets with 0.9 and 0.1 fractions [b], resulting in a training and validation sizes of 1476 and 164 configurations, respectively. The resulting training set was then increased by adding ten evenly spaced frames from the relaxation of each *pristine* host structure (110 configurations) — as this was observed to improve performance (Table 14 and Figs. 21 and 22).

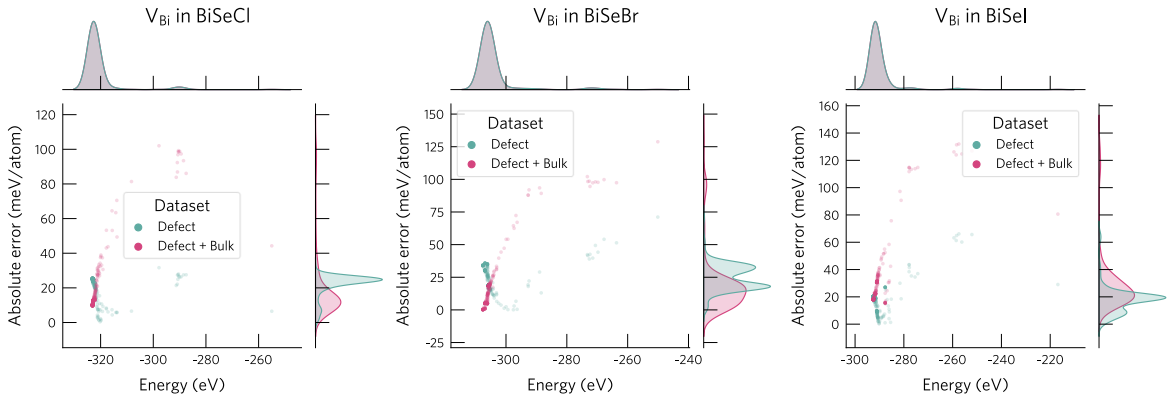


Supplementary Figure 20: Plot of the first two principal components of the feature space for the pristine structures, showing the selection of the chalcohalide systems. The structures were encoded with the SOAP descriptor, whose dimensions were reduced for plotting with Principal Component Analysis.[90].

Supplementary Table 14: Mean absolute errors, root mean squared errors and Spearman coefficients on the test set for the chalcohalide model when only training on defect data and when training on both defect and bulk data. The distribution of the absolute errors is shown in Figs. 21 and 22.

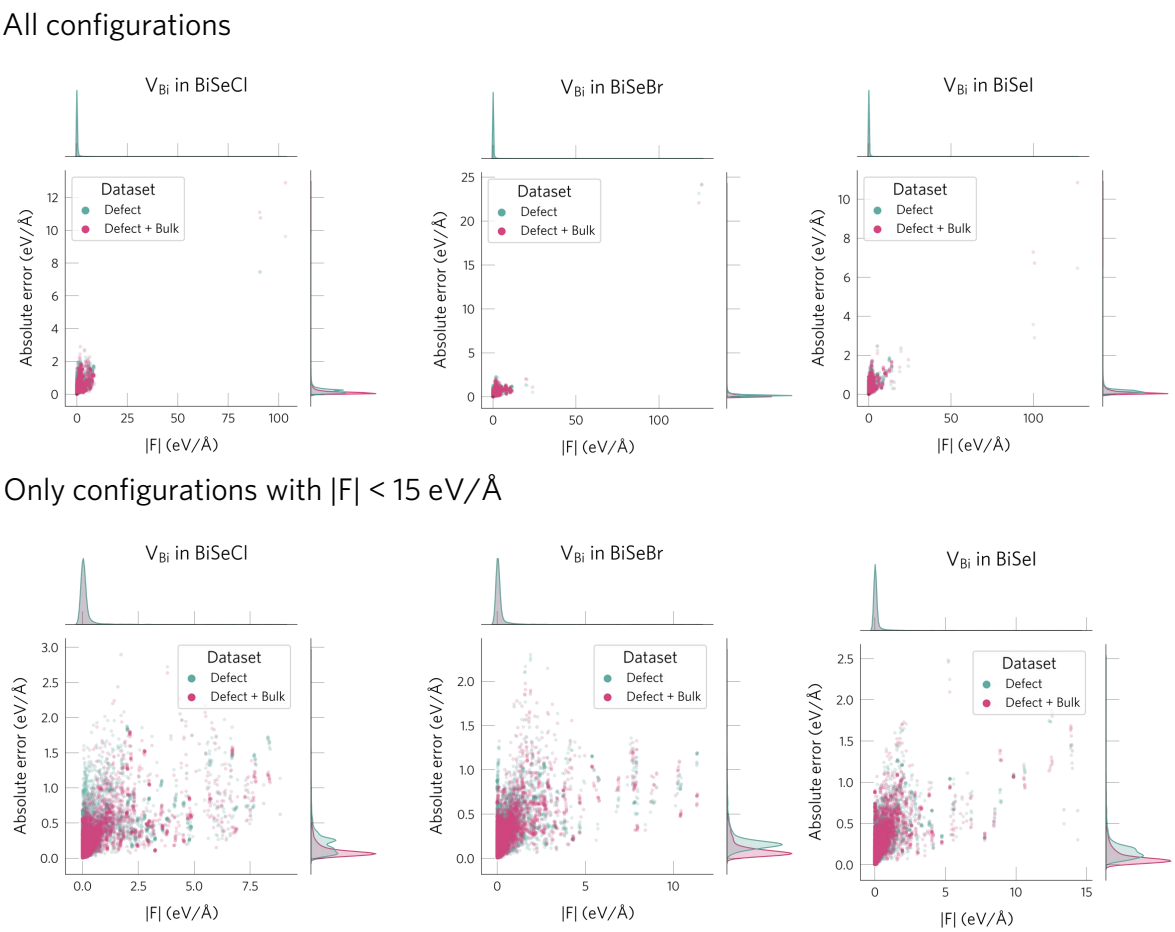
Dataset	MAE_E (meV/atom)	$RMSE_E$ (meV/atom)	ρ	MAE_F (meV/Å)	$RMSE_F$ (meV/Å)	MAE_S (GPa)	$RMSE_S$ (GPa)
Bulk + Defect	21.9	30.6	0.82	63.8	153.7	0.10	0.17
Defect only	21.3	23.1	0.72	97.6	175.2	0.10	0.16

[b] This split was done by selecting evenly spaced frames for the validation data, to ensure that both sets are representative of the original dataset.



Supplementary Figure 21: Comparison of absolute energy error distributions for the test systems when training on defect (green) and both defect and bulk data (pink). As seen for BiSeCl and BiSeBr, adding bulk data to the training set reduces the errors for the low-energy configurations.

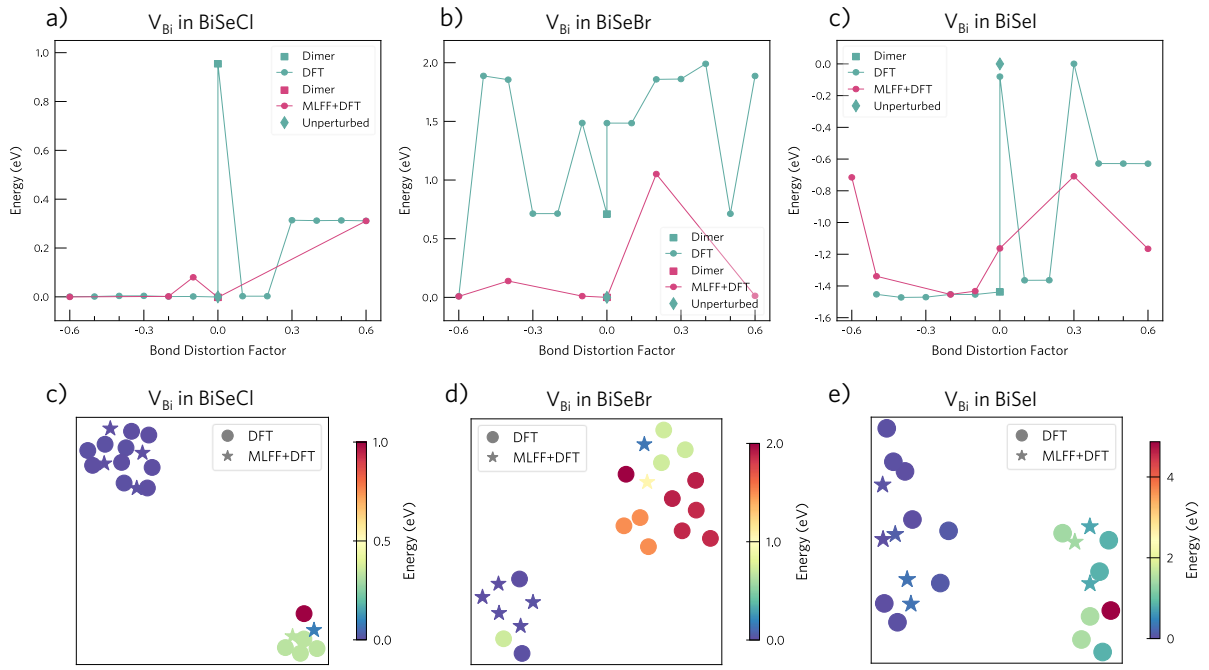
As shown in Table 15, the mean absolute errors are slightly lower than for the full model (trained on all compositions), confirming that smaller datasets can be used when targeting more similar host structures since their PES is easier to learn. After applying our MLFF+DFT approach to the test systems, it identifies the correct ground state for all three defects, while reducing the number of DFT calculations by 53%. Further, we note that the candidate structures selected in our approach (by relaxing the initial sampling structures with the MLFF and then selecting the structures with a unique SOAP fingerprint[90] for the defect site) target the low-energy region of the PES, as demonstrated in Fig. 23. Further, it also identifies two low-energy metastable structures that are missed with the DFT-only approach, validating that the model learns to suggest good candidate structures.



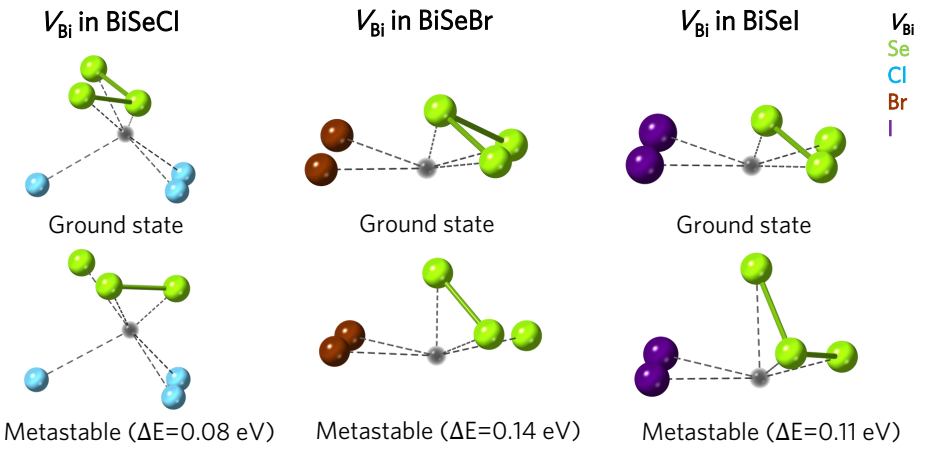
Supplementary Figure 22: Comparison of absolute force error distributions for the test systems when training on defect (green) and both defect and bulk data (pink). a) All configurations and b) removing outliers with high forces ($|F| < 15$ eV/Å). Adding bulk data to the training set reduces the force errors for the low-energy configurations.

Supplementary Table 15: Mean absolute errors and Spearman coefficients (ρ) for training, validation and test sets of the chalcohalide model (training on both defect and bulk data).

Set	E (meV/atom)	F (meV/Å)	S (GPa)	ρ
Train	19.2	51.0	0.06	0.92
Val	13.3	51.7	0.05	0.91
Test	21.9	63.7	0.10	0.82



Supplementary Figure 23: Comparison of DFT and MLFF+DFT approaches. (a, b, c) Plots of final (DFT) energy versus initial distortion for $V_{\mathrm{Bi}}{}^{0}$ in BiSeCl, BiSeBr and BiSeI. The full DFT approach (green) is compared with our MLFF+DFT strategy (pink). Note that the MLFF+DFT successfully targets the low-energy regions of the PES. Further, we note that several low-energy metastable configurations are missed with the DFT-only approach but identified with the MLFF+DFT one. The label "Unperturbed" denotes the configuration obtained by relaxing the high symmetry ideal structure while the label "Dimer" denotes a targeted distortion that pushes two of the defect nearest neighbours towards each other. (c, d, e) 2D projection of structural similarity for final structures obtained with the full DFT (circles) and MLFF+DFT approach (stars), illustrating that the latter targets the low-energy regions of the PES (e.g. no red stars).



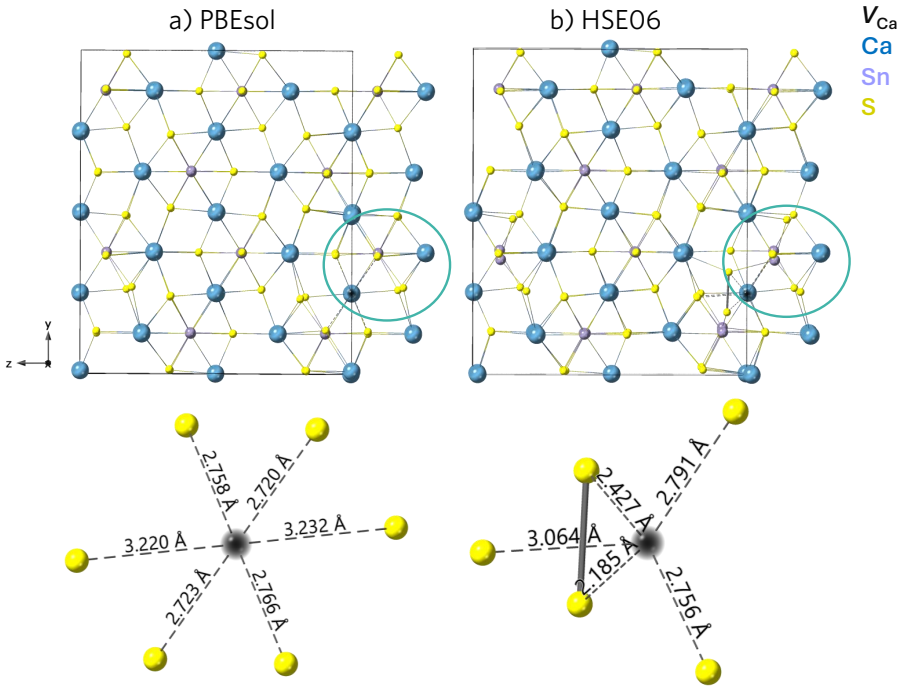
Supplementary Figure 24: Ground state structures for V_{Bi} in BiSeCl, BiSeBr and BiSeI. On the bottom, the low-energy metastable configurations that are only identified with the MLFF+DFT strategy are shown.

D. Structure searching with a semi-local functional

To demonstrate that defect structure searching should be performed with a hybrid functional rather than a semi-local functional, we compare the ground state structure obtained when performing the structure search with HSE06 and PBEsol for a set of representative defects from our test set that undergo energy-lowering reconstructions. As shown in Table 16, for 83% of defects there are significant differences between the ground state structures identified. The structural difference is quantified with the maximum distance between paired sites (d_{max}, in Å) and by analysing which anions form the anion-anion bonds in the structures. For $V_{\mathrm{Ca},0}$ in Ca_2SnS_4, the comparison is also shown visually in Fig. 25.

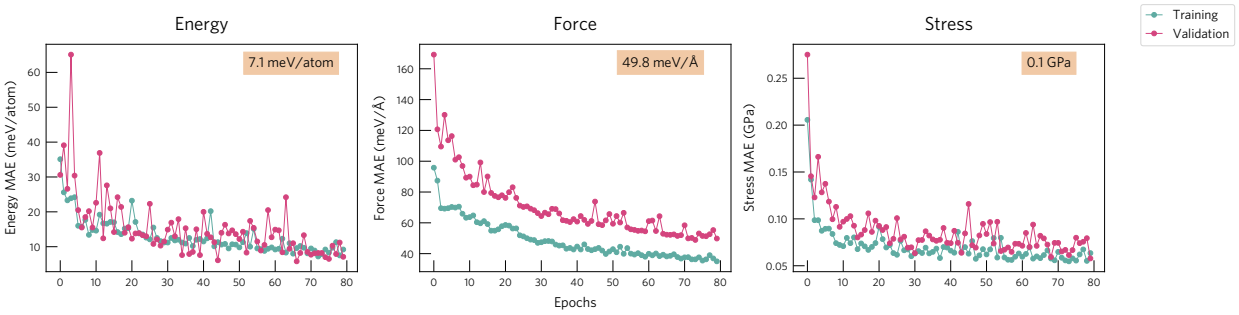
Supplementary Table 16: Comparison of the ground state structures identified with PBEsol and HSE06 for a set of representative defects. d_{max} denotes the maximum distance between paired sites (in Å) and ΔE indicates the energy difference between the ground state structure identified with HSE06 and PBEsol (evaluated with single-point HSE06 calculations on both geometries, i.e. $\Delta E = E_{\mathrm{gs,HSE06}} - E_{\mathrm{gs,PBEsol}}$). The columns 'Anion bonds (PBEsol)' and 'Anion bonds (HSE06)' list the atom indices of the anions forming the anion-anion bonds in the PBEsol and HSE06 structures and their corresponding bond distances, in Å. The results demonstrate that a semi-local functional tends to miss favourable defect structures due to its tendency to favour geometries with delocalised charges.

Hosts	Defects	d_{max}(Å)	ΔE(eV)	Anion bonds (PBEsol)	Anion bonds (HSE06)
Ca_2SnS_4	$V_{\mathrm{Ca},0}$	0.33	-0.42		S(53-77): 2.0
Ca_2SnS_4	$V_{\mathrm{Ca},4}$	0.03	-0.16	S(63-79): 2.1	S(63-79): 2.1
Ca_2SnS_4	$V_{\mathrm{Sn},8}$	0.37	-0.15	S(47-79): 2.8, S(65-79): 2.2, S(103-79): 2.1	S(47-103): 2.1, S(79-103): 2.1
$ZrSnS_3$	$V_{\mathrm{Zr},0}$	0.31	-0.14	S(95-151): 2.1, S(101-151): 2.1	S(77-151): 2.1, S(101-151): 2.1
$ZrSnS_3$	$V_{\mathrm{Sn},4}$	0.04	-0.12	S(95-153): 2.1	S(95-153): 2.1
BiSeCl	$V_{\mathrm{Bi},0}$	0.16	-0.12	Se(30-41): 2.5, Se(30-45): 2.5	Se(30-41): 2.3, Se(30-45): 3.0
BiSeBr	$V_{\mathrm{Bi},0}$	0.17	-0.25	Se(30-45): 2.5, Se(30-41): 2.4	Se(30-45): 2.3, Se(30-41): 3.0



Supplementary Figure 25: Comparison of the ground state structures obtained with PBEsol (a) and HSE06 (b) for $V_{\mathrm{Ca},0}$ in Ca_2SnS_4. The maximum distance between the paired sites of both structures is 0.33 Å, which is caused by the displacement of two of the defect's nearest neighbours to form a S-S bond in the HSE06 ground state. This dimerization reconstruction is not found with the PBEsol structure search, likely due to the tendency of semi-local functionals to favour delocalised charges (i.e. favouring the delocalisation of the two holes rather than their localisation in the S-S bond).

E. Extension to Alloys



Supplementary Figure 26: Evolution of training and validation errors with epoch number. The validation errors obtained for the epoch with the lowest validation loss are shown at the top right of each subplot.

Supplementary Table 17: Performance of the MLFF model on the $CdSe_xTe_{(1-x)}$ systems. For each inequivalent defect in each alloy, we show the number of local minima identified by the DFT-based (coarser) search and by the MLFF finer search. When a more favourable ground state (GS) configuration is identified by the MLFF search, the energy lowering and the reconstruction driving it are shown.

x	Defects	Num. local minima in DFT PES	Num. local minima in MLFF PES	Novel GS? (eV)	Reconstruction
0.2	V_{Cd_0}	2	5	-0.3	Se-Te → Te-Te
0.2	$V_{Cd_{10}}$	2	6	-0.5	Se-Se → Se-Te
0.2	V_{Cd_1}	2	4	-0.6	Se-Se → Te-Te
0.2	V_{Cd_4}	2	4	0.0	Te-Te → Te-Te
0.3	V_{Cd_0}	2	5	-0.4	Se-Te → Te-Te
0.3	$V_{Cd_{21}}$	3	5	-0.2	Se-Se → Se-Te
0.3	V_{Cd_2}	2	6	0.0	Se-Se → Se-Se
0.3	V_{Cd_4}	2	4	-0.2	Se-Se → Se-Te
0.3	V_{Cd_5}	2	4	0.0	Te-Te → Te-Te
0.5	V_{Cd_0}	1	6	-0.7	T_d[a] → Te-Te
0.5	V_{Cd_4}	1	5	-0.5	T_d → Te-Te
0.5	V_{Cd_1}	2	3	-0.5	Se-Te → Te-Te
0.5	V_{Cd_3}	2	4	-0.3	Se-Se → Se-Te
0.5	V_{Cd_5}	2	4	-0.4	Se-Se → Se-Se
0.6	V_{Cd_0}	2	4	-0.5	Se-Se → Se-Te
0.6	V_{Cd_2}	3	10	-0.1	Se-Se → Se-Te
0.6	V_{Cd_3}	2	5	0.0	Te-Te → Te-Te
0.6	V_{Cd_4}	3	3	0.0	Se-Se → Se-Se
0.8	V_{Cd_0}	3	5	-0.6	Se-Se → Se-Te
0.8	V_{Cd_1}	3	7	0.0	Se-Se → Se-Se
0.8	$V_{Cd_{25}}$	1	3	-0.6	T_d → Te-Te
0.8	V_{Cd_3}	3	4	-0.3	Se-Te → Te-Te
0.9	V_{Cd_0}	1	4	-0.7	T_d → Se-Te
0.9	V_{Cd_0}	3	6	0.0	Se-Se → Se-Se

[a] T_d denotes the tetrahedral (high-symmetry) coordination of the defect environment when no anion-anion bonds are formed.

References

[1]Sambur, J. & Brgoch, J. Unveiling the hidden influence of defects via experiment and data science. *Chem. Mater.* **35** (18), 7351–7354 (2023) .

[2]Shockley, W. & Read, W. T. Statistics of the Recombinations of Holes and Electrons. *Phys. Rev.* **87** (5), 835–842 (1952) .

[3]Kim, S., Márquez, J. A., Unold, T. & Walsh, A. Upper limit to the photovoltaic efficiency of imperfect crystals from first principles. *Energy Environ. Sci.* **13** (5), 1481–1491 (2020) .

[4]Maier, J. Thermodynamics of Electrochemical Lithium Storage. *Angew. Chem. Int. Ed.* **52** (19), 4998–5026 (2013) .

[5]Squires, A. G. *et al.* Low electronic conductivity of $Li_7La_3Zr_2O_{12}$ solid electrolytes from first principles. *Phys. Rev. Mater.* **6** (8), 085401 (2022) .

[6]Li, W. *et al.* Defect Engineering for Fuel-Cell Electrocatalysts. *Adv. Mater.* **32** (19), 1907879 (2020) .

[7]Pastor, E. *et al.* Electronic defects in metal oxide photocatalysts. *Nat. Rev. Mater.* **7** (7), 503–521 (2022) .

[8]Kehoe, A. B., Scanlon, D. O. & Watson, G. W. Role of lattice distortions in the oxygen storage capacity of divalently doped CeO_2. *Chem. Mater.* **23** (20), 4464–4468 (2011) .

[9]Ivády, V., Abrikosov, I. A. & Gali, A. First principles calculation of spin-related quantities for point defect qubit research. *npj Comput Mater* **4** (1), 76 (2018) .

[10]Weber, J. R. *et al.* Quantum computing with defects. *Proc. Natl. Acad. Sci. U.S.A.* **107** (19), 8513–8518 (2010) .

[11]Thomas, J. *et al.* A substitutional quantum defect in WS_2 discovered by high-throughput computational screening and fabricated by site-selective stm manipulation. Preprint at https://arxiv.org/abs/2309.08032 (2023).

[12]Dreyer, C. E., Alkauskas, A., Lyons, J. L., Janotti, A. & Van de Walle, C. G. First-principles calculations of point defects for quantum technologies. *Annu. Rev. Mater. Res.* **48** (1), 1–26 (2018) .

[13]Yan, Q., Kar, S., Chowdhury, S. & Bansil, A. The case for a defect genome initiative. *Adv Mater* 2303098 (2024) .

[14]Davidsson, J., Bertoldo, F., Thygesen, K. S. & Armiento, R. Absorption versus adsorp-

tion: in 2D materials. *npj 2D Mater. Appl.* **7** (1), 26 (2023) .

[15]Sluydts, M., Pieters, M., Vanhellemont, J., Speybroeck, V. V. & Cottenier, S. High-Throughput Screening of Extrinsic Point Defect Properties in Si and Ge: Database and Applications. *Chem. Mater.* **29** (3), 975–984 (2016) .

[16]Bertoldo, F., Ali, S., Manti, S. & Thygesen, K. S. Quantum point defects in 2d materials - the QPOD database. *npj Comput Mater* **8** (1), 56 (2022) .

[17]Huang, P. *et al.* Unveiling the complex structure-property correlation of defects in 2d materials based on high throughput datasets. *npj 2D Mater Appl* **7** (1), 1–10 (2023) .

[18]Medasani, B. *et al.* Predicting defect behavior in B2 intermetallics by merging ab initio modeling and machine learning. *npj Comput Mater* **2** (1), 1–10 (2016) .

[19]Rahman, M. H. *et al.* Accelerating defect predictions in semiconductors using graph neural networks. *APL Mach. Learn.* **2** (1), 016122 (2024) .

[20]Ivanov, V. *et al.* Database of semiconductor point-defect properties for applications in quantum technologies. Preprint at https://arxiv.org/abs/2303.16283 (2023).

[21]Kumagai, Y., Tsunoda, N., Takahashi, A. & Oba, F. Insights into oxygen vacancies from high-throughput first-principles calculations. *Phys. Rev. Mater.* **5**, 123803 (2021) .

[22]Deml, A. M., Holder, A. M., O'Hayre, R. P., Musgrave, C. B. & Stevanović, V. Intrinsic material properties dictating oxygen vacancy formation energetics in metal oxides. *J. Phys. Chem. Lett.* **6** (10), 1948–1953 (2015) .

[23]Broberg, D. *et al.* High-throughput calculations of charged point defect properties with semi-local density functional theory—performance benchmarks for materials screening applications. *npj Comput Mater* **9** (1), 72 (2023) .

[24]Mannodi-Kanakkithodi, A. *et al.* Universal machine learning framework for defect predictions in zinc blende semiconductors. *Patterns* **3** (3), 100450 (2022) .

[25]Varley, J. B., Samanta, A. & Lordi, V. Descriptor-based approach for the prediction of cation vacancy formation energies and transition levels. *J. Phys. Chem. Lett.* **8** (20), 5059–5063 (2017) .

[26]Wan, Z., Wang, Q.-D., Liu, D. & Liang, J. Data-driven machine learning model for the prediction of oxygen vacancy formation energy of metal oxide materials. *Phys. Chem. Chem. Phys.* **23** (29), 15675–15684 (2021) .

[27]Wexler, R. B., Gautam, G. S., Stechel, E. B. & Carter, E. A. Factors governing oxygen vacancy formation in oxide perovskites. *J. Am. Chem. Soc.* **143** (33), 13212–13227 (2021)

.

[28]Frey, N. C., Akinwande, D., Jariwala, D. & Shenoy, V. B. Machine Learning-Enabled Design of Point Defects in 2D Materials for Quantum and Neuromorphic Information Processing. *ACS Nano* **14** (10), 13406–13417 (2020) .

[29]Sharma, V., Kumar, P., Dev, P. & Pilania, G. Machine learning substitutional defect formation energies in ABO_3 perovskites. *J. Appl. Phys.* **128** (3), 034902 (2020) .

[30]Baldassarri, B. *et al.* Oxygen vacancy formation energy in metal oxides: High-throughput computational studies and machine-learning predictions. *Chem. Mater.* **35** (24), 10619–10634 (2023) .

[31]Park, S. *et al.* Exploring the latent chemical space of oxygen vacancy formation energy by a machine learning ensemble. *ACS Mater. Lett.* **0** (0), 66–72 (0) .

[32]Kazeev, N. *et al.* Sparse representation for machine learning the properties of defects in 2D materials. *npj Comput Mater* **9** (1), 113 (2023) .

[33]Choudhary, K. & Sumpter, B. G. Can a deep-learning model make fast predictions of vacancy formation in diverse materials? *AIP Adv.* **13** (9), 095109 (2023) .

[34]Zhao, X., Yu, S., Zheng, J., Reece, M. J. & Zhang, R.-Z. Machine learning of carbon vacancy formation energy in high-entropy carbides. *J. Eur. Ceram. Soc.* **43** (4), 1315–1321 (2023) .

[35]Manzoor, A. *et al.* Machine learning based methodology to predict point defect energies in multi-principal element alloys. *Front. Mater.* **8**, 673574 (2021) .

[36]Polak, M. P., Jacobs, R., Mannodi-Kanakkithodi, A., Chan, M. K. Y. & Morgan, D. Machine learning for impurity charge-state transition levels in semiconductors from elemental properties using multi-fidelity datasets. *J. Chem. Phys.* **156** (11), 114110 (2022) .

[37]Witman, M. D., Goyal, A., Ogitsu, T., McDaniel, A. H. & Lany, S. Defect graph neural networks for materials discovery in high-temperature clean-energy applications. *Nat Comput Sci* **3** (8), 675–686 (2023) .

[38]Arrigoni, M. & Madsen, G. K. H. Evolutionary computing and machine learning for discovering of low-energy defect configurations. *npj Comput. Mater.* **7** (71), 1–13 (2021) .

[39]Kavanagh, S. R., Walsh, A. & Scanlon, D. O. Rapid Recombination by Cadmium Vacancies in CdTe. *ACS Energy Lett.* **6** (4), 1392–1398 (2021) .

[40]Mosquera-Lois, I. & Kavanagh, S. R. In search of hidden defects. *Matter* **4** (8), 2602–2605

(2021) .

[41]Mosquera-Lois, I., Kavanagh, S. R., Walsh, A. & Scanlon, D. O. Identifying the ground state structures of point defects in solids. *npj Comput. Mater.* **9**, 1–11 (2023) .

[42]Wang, X., Kavanagh, S. R., Scanlon, D. O. & Walsh, A. Four-electron negative-U vacancy defects in antimony selenide. *Phys. Rev. B* **108**, 134102 (2023) .

[43]Wang, X., Kavanagh, S. R., Scanlon, D. O. & Walsh, A. Upper efficiency limit of Sb_2Se_3 solar cells. Preprint at https://arxiv.org/abs/2402.04434 (2024).

[44]Morris, A. J., Pickard, C. J. & Needs, R. J. Hydrogen/nitrogen/oxygen defect complexes in silicon from computational searches. *Phys. Rev. B* **80**, 144112 (2009) .

[45]Mulroue, J., Morris, A. J. & Duffy, D. M. Ab initio study of intrinsic defects in zirconolite. *Phys. Rev. B* **84**, 094118 (2011) .

[46]Al-Mushadani, O. K. & Needs, R. J. Free-energy calculations of intrinsic point defects in silicon. *Phys. Rev. B* **68** (23), 235205 (2003) .

[47]Kononov, A., Lee, C.-W., Shapera, E. & Schleife, A. Identifying native point defect configurations in α-alumina. *J. Phys. Condens.* **35** (33), 334002 (2023) .

[48]Schaarschmidt, M. *et al.* Learned force fields are ready for ground state catalyst discovery. Preprint at https://arxiv.org/abs/2209.12466 (2022).

[49]Lan, J. *et al.* AdsorbML: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials. *npj Comput Mater* **9** (1), 172 (2023) .

[50]Heinen, S., von Rudorff, G. F. & von Lilienfeld, O. A. Transition state search and geometry relaxation throughout chemical compound space with quantum machine learning. *J. Chem. Phys.* **157** (22), 221102 (2022) .

[51]Lany, S. & Zunger, A. Metal-Dimer Atomic Reconstruction Leading to Deep Donor States of the Anion Vacancy in II-VI and Chalcopyrite Semiconductors. *Phys. Rev. Lett.* **93** (15), 156404 (2004) .

[52]Kang, J. & Wang, L.-W. High defect tolerance in lead halide perovskite $CsPbBr_3$. *J. Phys. Chem. Lett.* **8** (2), 489–493 (2017) .

[53]Wilson, D. J., Sokol, A. A., French, S. A. & Catlow, C. R. A. Defect structures in the silver halides. *Phys. Rev. B* **77** (6), 064115 (2008) .

[54]Zhao, Y. *et al.* Correlations between Immobilizing Ions and Suppressing Hysteresis in Perovskite Solar Cells. *ACS Energy Lett.* **1** (1), 266–272 (2016) .

[55]Ágoston, P., Erhart, P., Klein, A. & Albe, K. Geometry, electronic structure and ther-

modynamic stability of intrinsic point defects in indium oxide. *J. Phys. Condens. Matter* **21** (45), 455801 (2009) .

[56]Han, D., Du, M.-H., Dai, C.-M., Sun, D. & Chen, S. Influence of defects and dopants on the photovoltaic performance of Bi_2S_3: first-principles insights. *J. Mater. Chem. A* **5** (13), 6200–6210 (2017) .

[57]Meggiolaro, D., Ricciarelli, D., Alasmari, A. A., Alasmary, F. A. S. & De Angelis, F. Tin versus Lead Redox Chemistry Modulates Charge Trapping and Self-Doping in Tin/Lead Iodide Perovskites. *J. Phys. Chem. Lett.* **11** (9), 3546–3556 (2020) .

[58]Erhart, P., Klein, A. & Albe, K. First-principles study of the structure and stability of oxygen defects in zinc oxide. *Phys. Rev. B* **72** (8), 085213 (2005) .

[59]Sokol, A. A., Walsh, A. & Catlow, C. R. A. Oxygen interstitial structures in close-packed metal oxides. *Chem. Phys. Lett.* **492** (1), 44–48 (2010) .

[60]Evarestov, R. A., Jacobs, P. W. M. & Leko, A. V. Oxygen interstitials in magnesium oxide: A band-model study. *Phys. Rev. B* **54** (13), 8969–8972 (1996) .

[61]Kotomin, E. A. & Popov, A. I. Radiation-induced point defects in simple oxides. *Nucl. Instrum. Methods Phys. Res. B* **141** (1), 1–15 (1998) .

[62]Burbano, M., Scanlon, D. O. & Watson, G. W. Sources of Conductivity and Doping Limits in CdO from Hybrid Density Functional Theory. *J. Am. Chem. Soc.* **133** (38), 15065–15072 (2011) .

[63]Scanlon, D. O. & Watson, G. W. On the possibility of p-type SnO_2. *J. Mater. Chem.* **22** (48), 25236–25245 (2012) .

[64]Godinho, K. G., Walsh, A. & Watson, G. W. Energetic and Electronic Structure Analysis of Intrinsic Defects in SnO_2. *J. Phys. Chem. C* **113** (1), 439–448 (2009) .

[65]Scanlon, D. O. *et al.* Nature of the Band Gap and Origin of the Conductivity of PbO_2 Revealed by Theory and Experiment. *Phys. Rev. Lett.* **107** (24), 246402 (2011) .

[66]Keating, P. R. L., Scanlon, D. O., Morgan, B. J., Galea, N. M. & Watson, G. W. Analysis of Intrinsic Defects in CeO2 Using a Koopmans-Like GGA+U Approach. *J. Phys. Chem. C* **116** (3), 2443–2452 (2012) .

[67]Walsh, A., Da Silva, J. L. F. & Wei, S.-H. Interplay between Order and Disorder in the High Performance of Amorphous Transparent Conducting Oxides. *Chem. Mater.* **21** (21), 5119–5124 (2009) .

[68]Whalley, L. D., Crespo-Otero, R. & Walsh, A. H-Center and V-Center Defects in Hybrid

Halide Perovskites. *ACS Energy Lett.* **2** (12), 2713–2714 (2017) .

[69]Agiorgousis, M. L., Sun, Y.-Y., Zeng, H. & Zhang, S. Strong Covalency-Induced Recombination Centers in Perovskite Solar Cell Material CH3NH3PbI3. *J. Am. Chem. Soc.* **136** (41), 14570–14575 (2014) .

[70]Whalley, L. D. *et al.* Giant Huang–Rhys Factor for Electron Capture by the Iodine Intersitial in Perovskite Solar Cells. *J. Am. Chem. Soc.* **143** (24), 9123–9128 (2021) .

[71]Motti, S. G. *et al.* Defect Activity in Lead Halide Perovskites. *Adv Mater* **31** (47), 1901183 (2019) .

[72]Xiao, Z., Meng, W., Wang, J. & Yan, Y. Defect properties of the two-dimensional (CH3NH3)2Pb(SCN)2I2 perovskite: a density-functional theory study. *Phys. Chem. Chem. Phys.* **18** (37), 25786–25790 (2016) .

[73]Na-Phattalung, S. *et al.* First-principles study of native defects in anatase TiO_2. *Phys. Rev. B* **73**, 125205 (2006) .

[74]Li, K., Willis, J., Kavanagh, S. R. & Scanlon, D. O. Computational prediction of an antimony-based n-type transparent conducting oxide: F-doped Sb_2O_5. *Chem. Mater.* **36** (6), 2907–2916 (2024) .

[75]Scanlon, D. O. Defect engineering of $basno_3$ for high-performance transparent conducting oxide applications. *Phys. Rev. B* **87**, 161201 (2013) .

[76]Cen, J., Zhu, B., Kavanagh, S. R., Squires, A. G. & Scanlon, D. O. Cation disorder dominates the defect chemistry of high-voltage LiMn1.5Ni0.5O4 (LMNO) spinel cathodes. *J. Mater. Chem. A* **11** (25), 13353–13370 (2023) .

[77]Mosquera-Lois, I., Kavanagh, S. R., Walsh, A. & Scanlon, D. O. ShakeNBreak: Navigating the defect configurational landscape. *J. Open Source Softw.* **7**, 4817 (2022) .

[78]NIST Chemistry WebBook. https://doi.org/10.18434/M32147. (accessed May 2023).

[79]Qi, J., Ko, T. W., Wood, B. C., Pham, T. A. & Ong, S. P. Robust training of machine learning interatomic potentials with dimensionality reduction and stratified sampling. *npj Comput. Mater.* **10**, 1–11 (2024) .

[80]Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat Comput Sci* **2** (11), 718–728 (2022) .

[81]Deng, B. *et al.* Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Mach. Intell.* **5** (9), 1031–1041 (2023) .

[82]Merchant, A. *et al.* Scaling deep learning for materials discovery. *Nature* **624** (7990),

80–85 (2023) .

[83]Batatia, I. *et al.* A foundation model for atomistic materials chemistry. Preprint at https://arxiv.org/abs/2401.00096 (2024).

[84]Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csanyi, G. Koyejo, S. *et al.* (eds) *MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields*. (eds Koyejo, S. *et al.*) *Advances in Neural Information Processing Systems*, Vol. 35, 11423–11436 (2022).

[85]Chen, C. & Ong, S. P. M3GNet (version 0.2.4), 2023.

[86]Salzbrenner, P. T. *et al.* Developments and further applications of ephemeral data derived potentials. *J. Chem. Phys.* **159** (14), 144801 (2023) .

[87]Pickard, C. J. Ephemeral data derived potentials for random structure search. *Phys. Rev. B* **106** (1), 014102 (2022) .

[88]Musielewicz, J., Wang, X., Tian, T. & Ulissi, Z. FINETUNA: fine-tuning accelerated molecular simulations. *Mach. Learn. Technol.* **3** (3), 03LT01 (2022) .

[89]Jung, H., Sauerland, L., Stocker, S., Reuter, K. & Margraf, J. T. Machine-learning driven global optimization of surface adsorbate geometries. *npj Comput Mater* **9** (1), 114 (2023) .

[90]Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013) .

[91]Hu, Y.-J. First-principles approaches and models for crystal defect energetics in metallic alloys. *Comput. Mater. Sci.* **216**, 111831 (2023) .

[92]Piochaud, J. B. *et al.* First-principles study of point defects in an fcc Fe-10Ni-20Cr model alloy. *Phys. Rev. B* **89**, 024101 (2014) .

[93]Guan, H. *et al.* Chemical environment and magnetic moment effects on point defect formations in CoCrNi-based concentrated solid-solution alloys. *Acta Mater.* **187**, 122–134 (2020) .

[94]Rio, E. d. *et al.* Formation energy of vacancies in FeCr alloys: Dependence on Cr concentration. *J. Nucl. Mater.* **408** (1), 18–24 (2011) .

[95]Zhang, Y. *et al.* Influence of chemical disorder on energy dissipation and defect evolution in concentrated solid solution alloys. *Nat Commun* **6** (1), 8736 (2015) .

[96]Zhang, Y. *et al.* Atomic-level heterogeneity and defect dynamics in concentrated solid-solution alloys. *Curr. Opin. Solid State Mater. Sci.* **21** (5), 221–237 (2017) .

[97]Arora, G., Bonny, G., Castin, N. & Aidhy, D. S. Effect of different point-defect energetics in $Ni_{80}X_{20}$ (X=Fe, Pd) on contrasting vacancy cluster formation from atomistic simulations. *Acta Mater* **15**, 100974 (2021) .

[98]Zhao, S., Stocks, G. M. & Zhang, Y. Defect energetics of concentrated solid-solution alloys from ab initio calculations: Ni0.5Co0.5, Ni0.5Fe0.5, Ni0.8Fe0.2 and Ni0.8Cr0.2. *Phys. Chem. Chem. Phys.* **18** (34), 24043–24056 (2016) .

[99]Manzoor, A. & Zhang, Y. Influence of Defect Thermodynamics on Self-Diffusion in Complex Concentrated Alloys with Chemical Ordering. *JOM* **74** (11), 4107–4120 (2022) .

[100]Zhao, S., Egami, T., Stocks, G. M. & Zhang, Y. Effect of d electrons on defect properties in equiatomic NiCoCr and NiCoFeCr concentrated solid solution alloys. *Phys. Rev. Mater.* **2** (1), 013602 (2018) .

[101]Li, C. *et al.* First principle study of magnetism and vacancy energetics in a near equimolar NiFeMnCr high entropy alloy. *J. Appl. Phys* **125** (15), 155103 (2019) .

[102]Manzoor, A., Zhang, Y. & Aidhy, D. S. Factors affecting the vacancy formation energy in Fe70Ni10Cr20 random concentrated alloy. *Comput. Mater. Sci.* **198**, 110669 (2021) .

[103]Muzyk, M., Nguyen-Manh, D., Kurzydłowski, K. J., Baluc, N. L. & Dudarev, S. L. Phase stability, point defects, and elastic properties of W-V and W-Ta alloys. *Phys. Rev. B* **84**, 104115 (2011) .

[104]Wang, Y. *et al.* Cation disorder engineering yields $AgBiS_2$ nanocrystals with enhanced optical absorption for efficient ultrathin solar cells. *Nat. Photon.* **16** (3), 235–241 (2022) .

[105]Williford, R., Weber, W., Devanathan, R. & Gale, J. Effects of Cation Disorder on Oxygen Vacancy Migration in $Gd_2Ti_2O_7$. *J. Electroceramics* **3** (4), 409–424 (1999) .

[106]Quadir, S. *et al.* Short- and Long-Range Cation Disorder in $(Ag_xCu_{1-x})_2ZnSnSe_4$ Kesterites. *Chem. Mater.* **34** (15), 7058–7068 (2022) .

[107]Morrow, J. D. *et al.* Understanding Defects in Amorphous Silicon with Million-Atom Simulations and Machine Learning. *Angew. Chem. Int. Ed.* e202403842 .

[108]Riebesell, J. *et al.* Matbench discovery – an evaluation framework for machine learning crystal stability prediction. Preprint at https://arxiv.org/html/2308.14920v2 (2023).

[109]Shimizu, K. *et al.* Using neural network potentials to study defect formation and phonon properties of nitrogen vacancies with multiple charge states in GaN. *Phys. Rev. B* **106** (5), 054108 (2022) .

[110]Ko, T. W., Finkler, J. A., Goedecker, S. & Behler, J. General-Purpose Machine Learning Potentials Capturing Nonlocal Charge Transfer. *Acc. Chem. Res.* **54** (4), 808–817 (2021) .

[111]Kavanagh, S. R., Scanlon, D. O., Walsh, A. & Freysoldt, C. Impact of metastable defect structures on carrier recombination in solar cells. *Faraday Discuss.* **239**, 339–356 (2022) .

[112]Mosquera-Lois, I., Kavanagh, S. R., Klarbring, J., Tolborg, K. & Walsh, A. Imperfections are not 0 K: free energy of point defects in crystals. *Chem. Soc. Rev.* **52** (17), 5812–5826 (2023) .

[113]Pols, M., Brouwers, V., Calero, S. & Tao, S. How fast do defects migrate in halide perovskites: insights from on-the-fly machine-learned force fields. *Chem. Commun.* **59** (31), 4660–4663 (2023) .

[114]Freysoldt, C. *et al.* First-principles calculations for point defects in solids. *Rev. Mod. Phys.* **86**, 253–305 (2014) .

[115]Lany, S. & Zunger, A. Assessment of correction methods for the band-gap problem and for finite-size effects in supercell defect calculations: Case studies for ZnO and GaAs. *Phys. Rev. B* **78**, 235104 (2008) .

[116]Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened coulomb potential. *J. Chem. Phys.* **118** (18), 8207–8215 (2003) .

[117]Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6** (1), 15–50 (1996) .

[118]Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993) .

[119]Kresse, G. & Hafner, J. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B* **49**, 14251–14269 (1994) .

[120]Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Comput. Mater. Sci.* **111**, 218–230 (2016) .

[121]Uhrin, M., Huber, S. P., Yu, J., Marzari, N. & Pizzi, G. Workflows in AiiDA: Engineering a high-throughput, event-based engine for robust and modular computational workflows. *Comput. Mater. Sci.* **187**, 110086 (2021) .

[122]Huber, S. P. *et al.* AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Sci. Data* **7** (1), 300 (2020) .

[123]Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013) .

[124]Jain, A. *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials* **1** (1), 011002 (2013) .

[125]Ong, S. P. *et al.* The materials application programming interface (API): A simple, flexible and efficient API for materials data based on REpresentational state transfer (REST) principles. *Comput. Mater. Sci.* **97**, 209–215 (2015) .

[126]Shen, J.-X. & Varley, J. pymatgen-analysis-defects: A python package for analyzing point defects in crystalline materials. *J. Open Source Softw.* **9** (93), 5941 (2024) .

[127]Shen, J.-X., Voss, L. F. & Varley, J. B. Simulating charged defects at database scale. *J. Appl. Phys.* **135** (14), 145102 (2024) .

[128]Larsen, A. H. *et al.* The atomic simulation environment—a python library for working with atoms. *J. Condens. Matter Phys.* **29** (27), 273002 (2017) .

[129]Kavanagh, S. R. *et al.* doped: Python toolkit for robust and repeatable charged defect supercell calculations. *J. Open Source Softw.* **9** (96), 6433 (2024) .

[130]Bitzek, E., Koskinen, P., Gähler, F., Moseler, M. & Gumbsch, P. Structural relaxation made simple. *Phys. Rev. Lett.* **97**, 170201 (2006) .

[131]van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008) .

[132]Hinton, G. E. & Roweis, S. Stochastic neighbor embedding. *Adv Neural Inf Process Syst* **15**, 857–864 (2002) .

[133]Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater* **31** (9), 3564–3572 (2019) .