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

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.

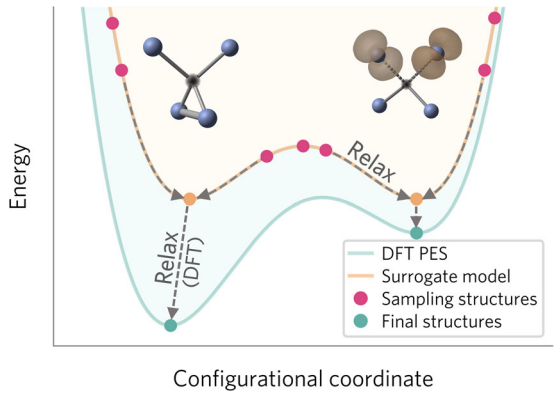

Fig. 1 | Schematic of a machine-learning surrogate model used to accelerate defect structure searching. The computationally efficient model learns the plausible defect reconstructions (local minima in the potential energy surface) and thus reduces the number of candidate structures relaxed with expensive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

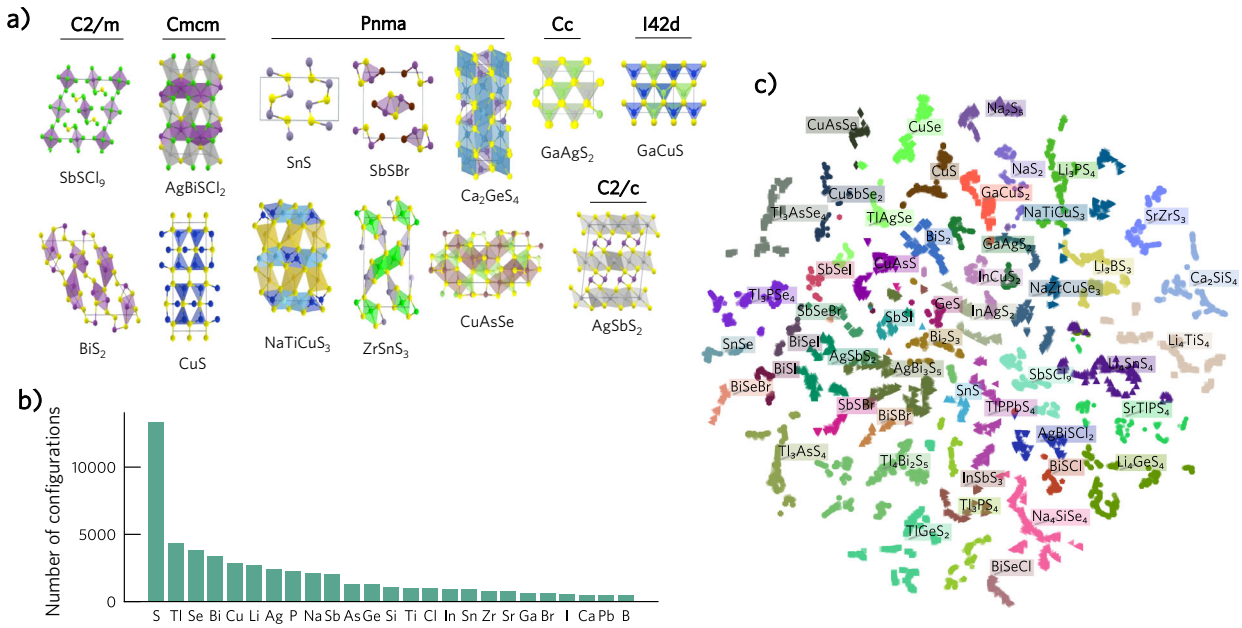

Fig. 2 | Distribution of the defect dataset generated with first-principles calculations. a Example host structures and their respective space groups. b) Number of configurations containing each element. c) Two-dimensional projection of structural similarity for defect configurations. Each configuration 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.

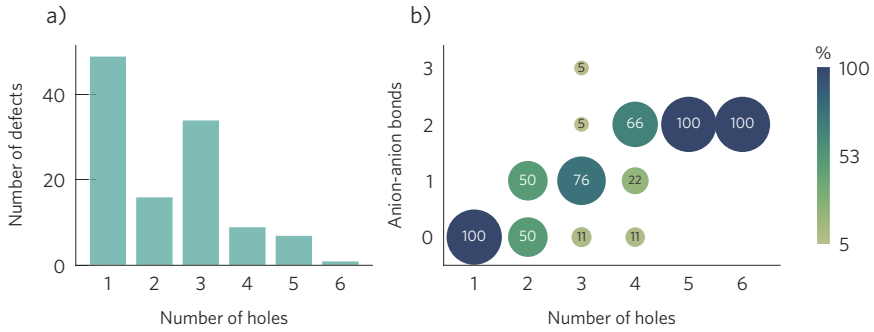

Fig. 3 | Analysis of the point defect dataset. **a** Distribution of number of holes produced upon defect formation. **b** Correlation between the number of anion—anion bonds formed and the number of holes created per defect. The label, colour, and size of the circles indicate the percentage of defects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

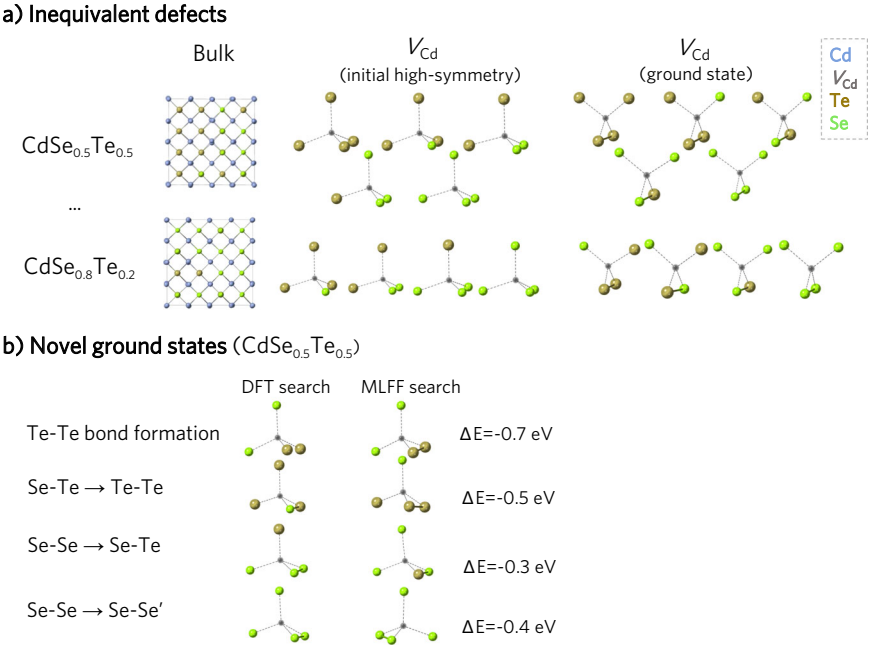

Fig. 4 | Structures for V_{Cd} in the $CdSe_xTe_{1-x}$ alloys. a Inequivalent defect environments for two of the $CdSe_xTe_{1-x}$ alloys ($x = 0.5, 0.8$). **b** Examples of the ground state configurations *only* identified through the finer MLFF+DFT search. These reconstructions are driven by either 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.