ARTICLE OPEN

Machine-learned impurity level prediction for semiconductors: the example of Cd-based chalcogenides

Arun Mannodi-Kanakkithodi[1]✉, Michael Y. Toriyama[1], Fatih G. Sen[1], Michael J. Davis[2], Robert F. Klie[3] and Maria K. Y. Chan[1]✉

The ability to predict the likelihood of impurity incorporation and their electronic energy levels in semiconductors is crucial for controlling its conductivity, and thus the semiconductor's performance in solar cells, photodiodes, and optoelectronics. The difficulty and expense of experimental and computational determination of impurity levels makes a data-driven machine learning approach appropriate. In this work, we show that a density functional theory-generated dataset of impurities in Cd-based chalcogenides CdTe, CdSe, and CdS can lead to accurate and generalizable predictive models of defect properties. By converting any semiconductor + impurity system into a set of numerical descriptors, regression models are developed for the impurity formation enthalpy and charge transition levels. These regression models can subsequently predict impurity properties in mixed anion CdX compounds (where X is a combination of Te, Se and S) fairly accurately, proving that although trained only on the end points, they are applicable to intermediate compositions. We make machine-learned predictions of the Fermi-level-dependent formation energies of hundreds of possible impurities in 5 chalcogenide compounds, and we suggest a list of impurities which can shift the equilibrium Fermi level in the semiconductor as determined by the dominant intrinsic defects. Machine learning predictions for the dominating impurities compare well with DFT predictions, revealing the power of machine-learned models in the quick screening of impurities likely to affect the optoelectronic behavior of semiconductors.

npj Computational Materials (2020)6:39 ; https://doi.org/10.1038/s41524-020-0296-7

INTRODUCTION

No crystalline material is devoid of defects and impurities. In fact, the imperfections in a crystal determine its properties as much as the regular arrangement of atoms do. When it comes to crystalline semiconducting materials, it is known that defects such as vacancies, native or impurity interstitials or substitutions, surface states, and grain boundaries can influence their optoelectronic properties. In the absence of external impurities, native defects determine the equilibrium Fermi level in the semiconductor, and thus the nature of conductivity (p-type, n-type or intrinsic) and charge carriers[1–3]. The introduction of impurity atoms can change the conductivity as determined by the dominant native defects, based on their formation enthalpies as a function of the Fermi energy[1,4]. Foresight about the impact of certain impurities on the electronic structure and conductivity of the material is crucial in either trying to curb their presence, or intentionally incorporating them in the semiconductor lattice to induce a desirable optoelectronic change.

It is important to be able to predict the electronic energy levels created by impurities in semiconductors. While shallow acceptor or donor levels are defined as defect levels close to the band edges and do not affect the recombination of charge carriers, deep defect levels can have both disastrous and potentially beneficial effects. Deep levels can act as non-radiative recombination centers for minority charge carriers, which significantly reduces their lifetime, impedes carrier collection or light emission, and drastically brings down the solar cell or photodiode efficiency and performance[5]. On the other hand, researchers have shown that in principle, energy levels in the band gap can be used as intermediate bands to facilitate absorption of sub-gap photons, which could enhance the absorption efficiencies[1,6,7].

Defect levels are often measured using methods like deep level transient spectroscopy (DLTS) and cathodoluminescence (CL)[8–11]. However, difficulties in incorporating specific impurities or dopants in a given compound and in attributing measured levels to specific defects make experimental methods less than ideal for an extensive study of defects and impurities in semiconductors. First-principles density functional theory (DFT) computations have been widely used instead to simulate substitutional or interstitial impurities and vacancies in crystalline materials using the supercell approach[12–14]. Impurity formation enthalpies, energy levels, and resulting absorption coefficients calculated from DFT typically match well with measured values[15–19]. However, DFT has limitations of its own: the requirement of large supercells, charge states, explicit image charge corrections, and an advanced level of theory (such as hybrid functionals[20] or *GW* corrections[21]) to accurately determine band gaps make these calculations generally expensive. Furthermore, prior knowledge is seldom utilized in informing or accelerating new defect calculations; there is an opportunity here for the creation of surrogate models based on previously generated data, such that impurity properties for fresh cases can be quickly and accurately estimated.

Today, machine learning (ML) has become an integral component of materials design[22]. Researchers have extracted models and design rules from materials data to drive the accelerated discovery of NiTi alloys for thermal hysteresis[23], design of polymer dielectrics for improved energy storage in capacitors[24,25], synthesis of new classes of compounds[26,27], identification of new and improved catalysts[28,29], and the design of experiments in a smart and 'adaptive' fashion[30]. ML-based design of materials usually begins with the generation of sufficient data for candidate materials in terms of a property ***P***, and the conversion of all materials in the chemical space into a

[1]Center for Nanoscale Materials, Argonne National Laboratory, Argonne, IL 60439, USA. [2]Chemical Sciences and Engineering Division, Argonne National Laboratory, Argonne, IL 60439, USA. [3]Department of Physics, University of Illinois at Chicago, Chicago, IL 60607, USA. ✉email: mannodiarun@anl.gov; mchan@anl.gov

unique numerical representation $\boldsymbol{X}$, referred to as descriptors, feature vectors, or fingerprints. This is followed by a mapping $\boldsymbol{X}\rightarrow\boldsymbol{P}$ between descriptors and properties using linear correlation[1,31] or a nonlinear regression technique such as ridge regression[32], support vector machine[33], random forest[34], LASSO (Least Absolute Shrinkage and Selection Operator)[35], or neural networks[36]. The result of such an approach is a trained predictive model which estimates $\boldsymbol{P}$ for any $\boldsymbol{X}$, with a statistical uncertainty or confidence interval that is also an output. The general outline for developing machine-learned predictive models for properties of materials based on DFT data and numerical descriptors is shown in Fig. 1a.

The prediction of defect or impurity formation enthalpies and energy levels can be accelerated by developing ML models trained from DFT data, as has been shown in the recent past[37]. As a demonstration of this approach, we take the example of Cd-based chalcogenides, which are important semiconductors for optoelectronic and solar cell applications[38–40], and apply ML algorithms on a dataset of DFT computed properties for hundreds of impurity types in CdTe, CdSe, and CdS. These compounds are chosen not only because CdTe-based cells are the second most commonly used photovoltaics after Si, but also because in recent years, significant improvements in the efficiency of CdTe solar cells have arisen due to the elimination of the CdS buffer layer and the introduction of Cd(Se,Te) into the absorber layer[41]. Therefore, the prediction of impurity levels in ternary Cd chalcogenides of various compositions is of technological importance. Each of these compounds, henceforth referred to as CdX (X = Te/Se/S), exists in the cubic Zinc Blende (ZB) structure[42] shown in Fig. 1b. Although delocalized states treated with PBE gives underestimated band-gaps[20,43], it has been shown that defect states at the PBE level can be accurately characterized and compared against experiments or higher levels of theory, e.g., as reflected in accurate defect transition levels that span the physical band gap[44]. It has also been shown in the past that suitable alignment schemes can be used to ensure DFT defect levels agree with values obtained from higher levels of theory[45]. In Fig. 1c, we plotted the band gaps computed from PBE and HSE06 functionals alongside the known, experimentally measured band gaps[46,47] for 5 compounds: CdTe, CdSe, CdS, and mixed anion compounds, $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$; the HSE06 computed values match well with experiments. Some discrepancies, eg. in the band gap of $CdTe_{0.5}Se_{0.5}$, could arise from the lack of structural relaxation performed in HSE, the neglect of spin-orbit coupling[48], or from the anion ordering not being adequately captured by SQS[49]. It was also reported that $CdTe_{1-x}Se_x$ compound is found in the Wurtzite rather than Zinc Blende structure, albeit only for $x > 0.6$[50]. It must be emphasized here that due to the high-throughput nature of this study and some limitations of the level of theory used, we accept uncertainties in computed band gaps and defect levels of up to 0.2 eV. The DFT computed lattice constants and band gaps for the 5 compounds are listed in Table SI-1.

In this work, we use both the PBE and HSE06 functionals to compute impurity properties in different CdX compounds; the eventual dataset of HSE impurity levels is one-fifth the size of the corresponding PBE dataset, owing to the 2 orders of magnitude difference in computational expense. We train separate ML models for impurity properties computed with PBE and HSE, and explore how models trained for lower fidelity (presumably, PBE) can inform the higher fidelity (presumably, HSE) predictions. We simulate impurities in several different defect sites in any CdX compound: one cation site (M_{Cd}, where M is the impurity atom), one or two anion sites (M_X) and three or four interstitial sites (M_i), based on whether it is a pure or mixed anion composition[51]; each of these sites have been pictured for CdTe in Fig. SI-1. Impurity atoms M are obtained by sweeping across the periodic table and selecting elements from periods II to VI, as shown in Fig. SI-2.

An outline of the work presented in this manuscript is shown in in Fig. 1d. DFT is used to compute the impurity formation enthalpy as a function of chemical potential (μ), charge (q) and Fermi energy (E_F), using Eq. (1) (in Methods), and the impurity charge transition levels using equation (2) (in Methods). ML models are trained for two types of properties: the neutral-state formation enthalpy ΔH ($E^f(q = 0)$) for Cd-rich to X-rich chemical potential

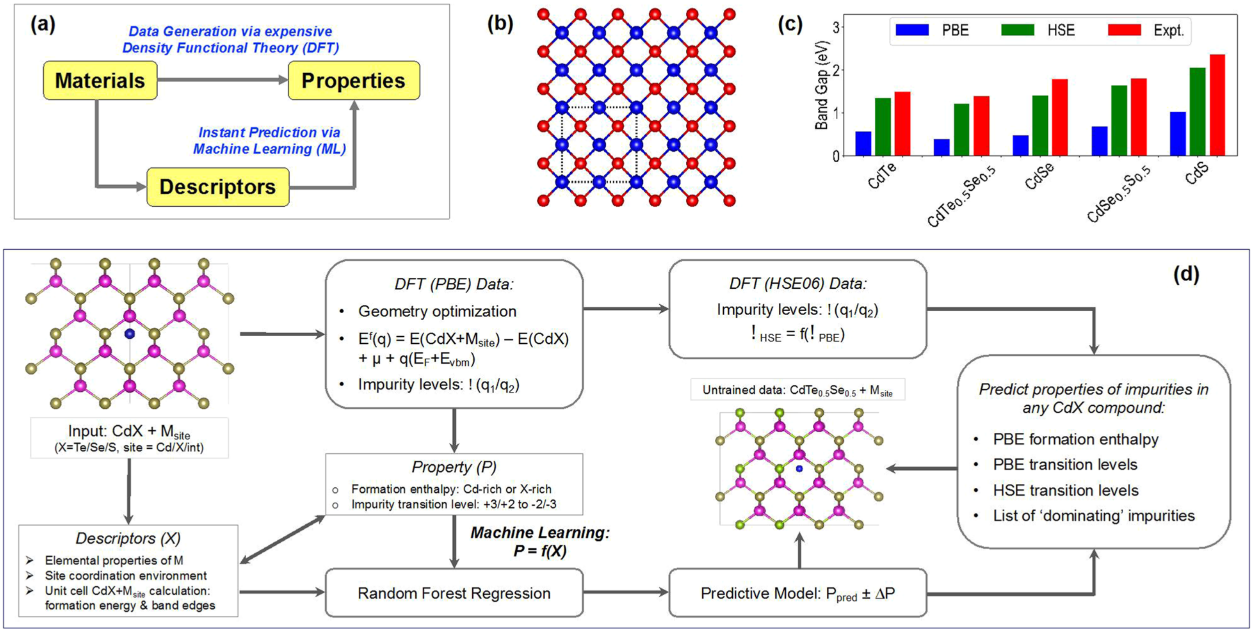
(a)
Data Generation via expensive Density Functional Theory (DFT)
Materials
Properties
Instant Prediction via Machine Learning (ML)
Descriptors
(b)
(c)
PBE
HSE
Expt.
Band Gap (eV)
0
1
2
3
CdTe
CdTe0.5Se0.5
CdSe
CdSe0.5S0.5
CdS
(d)
Input: CdX + M_site
(X=Te/Se/S, site = Cd/X/int)
DFT (PBE) Data:
Geometry optimization
E^f(q) = E(CdX+M_site) – E(CdX) + μ + q(E_F+E_vbm)
Impurity levels: ! (q1/q2)
DFT (HSE06) Data:
Impurity levels: ! (q1/q2)
! HSE = f(! PBE)
Untrained data: CdTe0.5Se0.5 + M_site
Predict properties of impurities in any CdX compound:
PBE formation enthalpy
PBE transition levels
HSE transition levels
List of 'dominating' impurities
Property (P)
Formation enthalpy: Cd-rich or X-rich
Impurity transition level: +3/+2 to -2/-3
Machine Learning:
P = f(X)
Descriptors (X)
Elemental properties of M
Site coordination environment
Unit cell CdX+M_site calculation: formation energy & band edges
Random Forest Regression
Predictive Model: P_pred ± ΔP

Fig. 1 Basic outline, structure and properties. **a** General outline of materials design process leading to ML-driven prediction of properties based on DFT data and intermediate step of converting materials to numerical descriptors. **b** The Zinc Blende structure adopted by CdTe, CdSe, and CdS. Cd atoms are shown in blue and Te/Se/S atoms in red. The unit cell has been indicated with dashed lines. **c** Comparison of band gaps computed at the PBE and HSE06 levels of theory with reported experimental values[46,47], for CdTe, CdSe, CdS, $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$. **d** Outline of the DFT and ML driven prediction of properties of impurities in Cd-based chalcogenides.

conditions), and various impurity charge transition levels, $\epsilon(q_1/q_2)$, which indicates the Fermi level at which the impurity containing system transitions from one stable charge state (q_1) to another (q_2). As shown in Fig. 1d, descriptors are generated for any CdX + M_{site} system (where M and 'site' refer to the impurity atom and defect site, respectively) based on tabulated elemental properties of M (such as ionic radii and electronegativity), site coordination environment, and properties computed from low-cost unit cell defect calculations. A regression algorithm is applied to map the descriptors to the properties, and predictive models are trained on the PBE formation enthalpy, PBE impurity transition levels, and HSE impurity transition levels. Comparisons in ML performance are made for different sets of descriptors, level of theory (PBE or HSE), and subset of computational data used for training. While models are trained for impurities in CdTe, CdSe, and CdS, we performed additional computations for selected impurities in $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$, to test the models' out-of-sample predictive ability. The power of this combined DFT + ML approach is illustrated with machine-learned predictions of Fermi level dependent formation enthalpies for the entire chemical space of impurities in CdTe, CdSe, CdS, $CdTe_{0.5}Se_{0.5}$, and $CdSe_{0.5}S_{0.5}$. These predictions, combined with the DFT computed formation enthalpies of intrinsic point defects (vacancies, anti-site, and interstitials) in each of the compounds, are used to obtain the list of impurities which can shift the equilibrium Fermi level (as determined by dominant native defects) and thus change the nature of conductivity in the semiconductor.

RESULTS AND DISCUSSION

PBE data: formation enthalpy and transition levels

The zero charge version of Eq. (1) (in Methods) was used to compute the formation enthalpy ΔH of impurities in CdX at Cd-rich and X-rich chemical potential conditions, for a few hundred impurity types. For CdTe, CdSe, and CdS, the neutral state impurity calculations are performed for each of the 63 elemental impurities as shown in Fig. SI-2, leading to a dataset of 315 ΔH ranges (Cd-rich to X-rich) for each compound. The chemical potential of any impurity atom is determined based on its stable compound with Te or its stable compound with Cd, referenced to its elemental standard state, where the structure for each compound or element is collected from the Materials Project[52]. The computed ΔH ranges have been plotted for the entire dataset in Figs. SI-3, SI-4, and SI-5, and for a few selected cases in Fig. 2. It can be seen from Fig. 2a–c that anti-site substitutional impurities such as Zr_{Te}, Se_{Cd}, and Na_S have high formation enthalpies and would be unstable, whereas other impurities like Ag_{Cd}, S_{Te}, and Br_S have much lower formation enthalpies. Further, 22 impurities were selected in $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$ across all 7 defect sites, and ΔH was computed for each to test the trained ML models (explained in the coming sections). The defect formation enthalpies for mixed anion compounds are shown in Fig. SI-6 and Fig. SI-7. A description of the PBE ΔH dataset across the 5 CdX compounds is provided in Table 1; data was generated for over 50% of the total chemical space of 1827 points.

Next, supercells containing impurity atoms were simulated in charge states of +3, +2, +1, −1, −2, and −3. For each of these calculations, the total DFT energies and charge correction terms (using Freysoldt's correction[14,53]) were obtained, and Eq. (2) (in Methods) was used to compute the various charge transition levels. All computed transition levels, namely, +3/+2, +2/+1, +1/0, 0/−1, −1/−2, and −2/−3, are plotted for the entire dataset of impurities in different sites in CdTe, CdSe, and CdS in Figs. SI-8, SI-9, and SI-10, respectively. This data has been presented once again for selected impurities in Fig. 2d–f. It should be noted that on occasion, transition levels like +1/−1 or +2/0 may exist, in which case the q/(q−1) and (q−1)/(q−2) transition levels are considered to be equal to the q/(q−2) transition level (for eg., +1/0 = 0/−1 = +1/−1). It can be seen from Fig. 2d–f that a number of impurities introduce energy levels in the band gap. This is attributed to the fact that an element prefers an oxidation state that is different from that of the element it is substituting; for instance, Bi_{Cd} leads to a net +1 charge in the system for a majority of the band gap and displays a +1/0 transition level close to the CBM, because of Bi adopting a +3 oxidimpurity transition levels.ation state as opposed to +2. Impurities that create mid-gap energy levels will be of interest if their formation enthalpies are low enough for them to be competitive with respect to dominant intrinsic point defects. Further, for the 22 additional impurities in $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$, all the transition levels are computed and plotted in Figs. SI-11 and SI-12. As listed in Table 1, DFT data was generated for 100% of the CdTe points, but 10% or less for CdSe, CdS, $CdTe_{0.5}Se_{0.5}$, and $CdSe_{0.5}S_{0.5}$. The total DFT dataset covers about 23% of the chemical space, providing a great opportunity for machine learning the remaining data points in a fraction of the time it takes to perform explicit DFT computations.

Descriptors for machine learning

As shown in Fig. 1a, the training of prediction models for material properties proceeds via the crucial intermediate step of descriptor generation. In this work, we utilize different sets of descriptors that represent the impurity atom and the defect site coordination, as well as some properties estimated from low-cost unit cell calculations. Similar descriptors were recently applied by us to represent impurities at the Pb-site in methylammonium lead bromide[1], from which we were able to train simple models to describe the formation enthalpy and charge transition levels. In a similar vein, we use the elemental properties of the impurity atom M, the number of Cd or X (Te/Se/S) neighboring atoms at the given defect site, and energetic and electronic properties calculated by modeling the M_{Cd}, M_X or M_i impurity in an 8-atom (Zinc Blende) CdX unit cell instead of a 64-atom supercell. The unit cell calculation is two orders of magnitude cheaper than the corresponding supercell calculation.

We apply different combinations of descriptors and use different regression algorithms to train predictive models for ΔH and $\epsilon(q_1/q_2)$. A base set of descriptors, namely the period and group of M, a defect site index (set as 0 for M_{Cd}, 1 for M_X, 0.50 for M_i(neutral site), 0.25 for M_i(Cd-site) and 0.75 for M_i(X-site)), and the number of Cd and X neighbors, is used in every combination. In addition, the elemental properties of M, such as the first ionization energy, electronegativity, and ionic radii, are used as descriptors to encode information about the structural and bonding characteristics of the impurity atom. Lastly, the impurity formation enthalpy at Cd-rich, intermediate, and X-rich chemical potential conditions, and the valence band and conduction band edges (universally aligned using the deep 5*s* semi-core state of Cd) calculated from the unit cell defect calculation are added as descriptors. Ultimately, we apply the following three sets of descriptors (in addition to the base set descriptors) independently to train the models:

1. Elemental properties
2. Unit cell defect properties
3. Elemental properties + unit cell defect properties

In Fig. 3a, we plot the Pearson correlation coefficient ($|r|$) between each descriptor and 9 different properties, namely the ΔH for Cd-rich, intermediate and X-rich conditions, and the +3/+2, +2/+1, +1/0, 0/−1, −1/−2, and −2/−3 impurity transition levels. It can be seen that while some of the elemental properties have a correlation of 0.40 to 0.50 with ΔH and $\epsilon(q_1/q_2)$, the unit cell defect properties exhibit the highest correlations. The valence and conduction band edges from unit cell defect calculations show a correlation of $|r|$ = 0.82 and $|r|$ = 0.74,

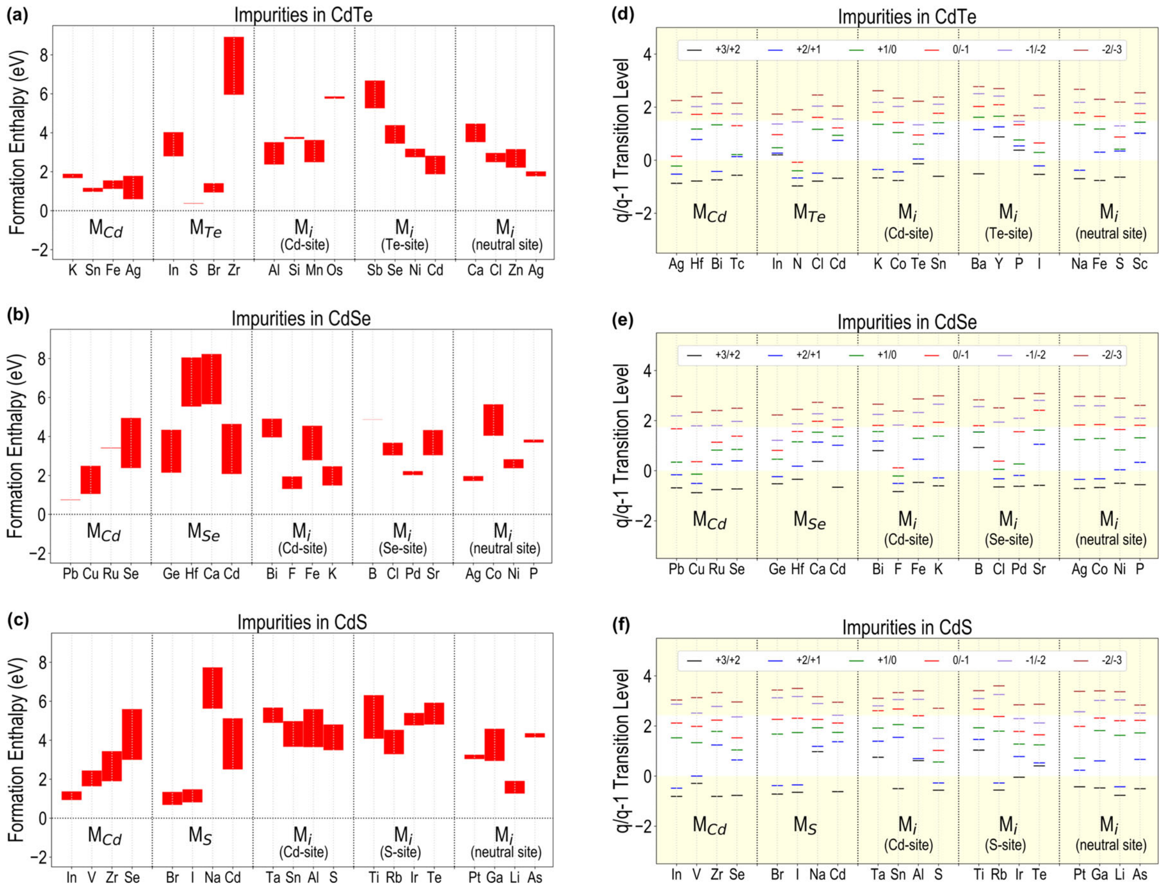

Fig. 2 PBE computed properties. Neutral-state impurity formation enthalpies computed at the PBE level of theory for selected impurity atoms in different sites in (**a**) CdTe, (**b**) CdSe and (**c**) CdS, and charge transition levels (from +3/+2 to −2/−3) calculated at the PBE level of theory for selected impurity atoms in different sites in (**d**) CdTe, (**e**) CdSe and (**f**) CdS. ΔH has been plotted for some very unstable impurities as well (like Hf_{Se} and Ta_i) to show the variety in the impurity property data that goes into training predictive models.

Table 1. Details of the DFT dataset.

Property	CdX	Impurity atoms	Defect sites	Transition levels	Total chemical space	DFT data	% of computed data
PBE ΔH	CdTe	63	5	–	63 × 5 = 315	315	100
	CdSe	63	5	–	63 × 5 = 315	315	100
	CdS	63	5	–	63 × 5 = 315	315	100
	$CdTe_{0.5}Se_{0.5}$	63	7	–	63 × 7 = 441	22	~5
	$CdSe_{0.5}S_{0.5}$	63	7	–	63 × 7 = 441	22	~5
	Total				1827	989	~54
PBE $\epsilon(q_1/q_2)$	CdTe	63	5	6	63 × 5 × 6 = 1890	1890	100
	CdSe	63	5	6	63 × 5 × 6 = 1890	198	~10.5
	CdS	63	5	6	63 × 5 × 6 = 1890	198	~10.5
	$CdTe_{0.5}Se_{0.5}$	63	7	6	63 × 7 × 6 = 2646	132	~7
	$CdSe_{0.5}S_{0.5}$	63	7	6	63 × 7 × 6 = 2646	132	~7
	Total				10962	2550	~23
HSE $\epsilon(q_1/q_2)$	CdTe	63	5	4	63 × 5 × 4 = 1260	240	~19
	CdSe	63	5	4	63 × 5 × 4 = 1260	132	~10.5
	CdS	63	5	4	63 × 5 × 4 = 1260	132	~10.5
	$CdTe_{0.5}Se_{0.5}$	63	7	4	63 × 7 × 4 = 1764	88	~5
	$CdSe_{0.5}S_{0.5}$	63	7	4	63 × 7 × 4 = 1764	88	~5
	Total				7308	680	~9.3

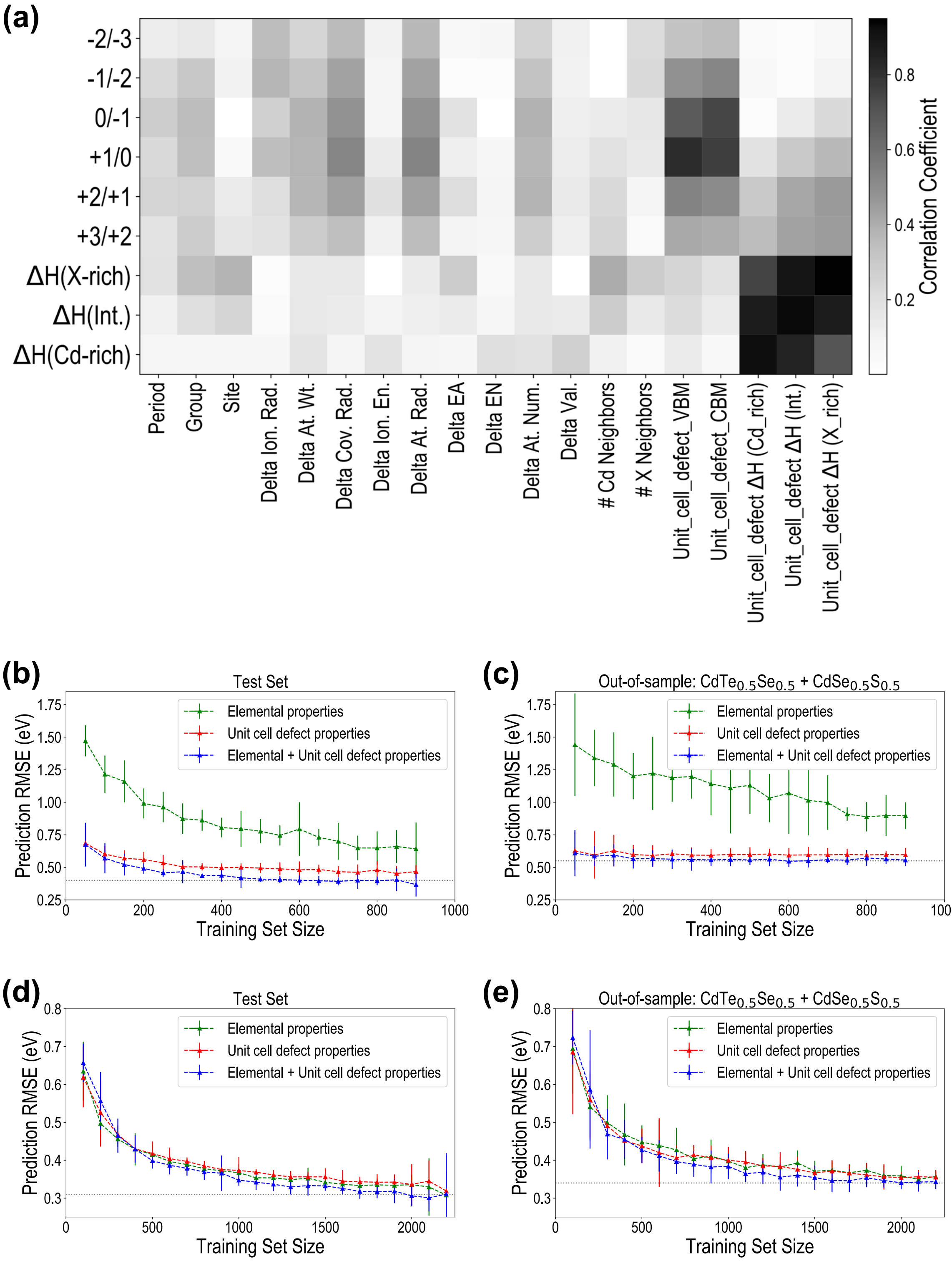

Fig. 3 Correlations and prediction errors. a Coefficient of linear correlation (|r|) between the properties of interest, ΔH and $\epsilon(q_1/q_2)$, and each of the descriptors. In **b** and **c**, prediction RMSE is plotted against the training set size for random forest models trained for ΔH (Cd-rich) using 3 different sets of features, for the test set points (total CdTe+CdSe+CdS dataset minus the training set) and the out-of-sample points (set of 22 impurities each in $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$) respectively. Similar plots are shown for $\epsilon(q_1/q_2)$ (at the PBE level of theory) (**d**) test set points and (**e**) out-of-sample points.

respectively, with the +1/0 and 0/−1 impurity transition levels. Further, ΔH (Cd-rich), ΔH (intermediate) and ΔH (X-rich) show a correlation of |r| > 0.90 with the corresponding ΔH values from unit cell defect calculations. When training predictive models for the impurity formation enthalpies and transition levels using these descriptors, one can expect more accurate predictions when including the unit cell defect properties as opposed to using elemental properties exclusively. However, while the unit cell defect calculations are not computationally intensive, the remaining descriptors can be generated with no additional computations at all, and thus have an advantage. In the next section, we examine the accuracy of regression models trained using different sets of descriptors.

Predictive models using regression

Three regression algorithms, namely Random Forest regression (RFR)[34], Kernel Ridge Regression (KRR)[32], and LASSO regression[35] (see details in Methods) were applied to train predictive models for ΔH and the ϵ(q/q−1) transition level for a given charge q. For each property, we trained models using the CdTe, CdSe, and CdS data, and data generated for $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$ was used to test the out-of-sample predictive power. For the impurity formation enthalpy, we train separate models for ΔH (Cd-rich) and ΔH (X-rich), since the two values provide the range of possible enthalpies over the chemical potential region of stability. The effects of training set size and choice of descriptors are studied by estimating the mean and standard deviation in prediction error over 100 different models trained (from different training sets) for any given case. For each regression technique, a grid-based search was applied to optimize the regression parameters, such as the number of trees and the number of features needed for splitting nodes in RFR, and the gaussian width and regularization parameter in KRR. To control overfitting of the ML models, *k*-fold cross-validation was used, wherein the training set is divided into *k* (here $k = 5$) sets and each of the *k* sets is used as an internal test set while training is performed using the remaining $k-1$ sets. The optimal ML parameters are obtained by minimizing the cross-validation error, that is the error on the *k*th set from the model trained using $k-1$ sets.

The root mean square errors (RMSE) of RFR models trained for ΔH (Cd-rich) are plotted as a function of the training set size for three sets of descriptors (each containing the base set), for the test set in Fig. 3b, and the out-of-sample points in Fig. 3c. All prediction errors steadily decrease with increasing training set size. It can be seen that for both the test and out-of-sample points, using just the elemental properties as descriptors leads to much higher errors than using the unit cell defect properties. The combination of elemental and unit cell defect properties shows the best prediction accuracies, and saturate fairly early to about 0.40 eV for the test set and 0.55 eV for the out-of-sample points, proving that reasonable prediction accuracies can be achieved with about 50% of the total data used for training. In Table 2, we have listed the RMSE for predictive models trained using RFR, KRR, and LASSO, with 90% of the dataset of CdTe, CdSe, and CdS points used for training, independently applying the three sets of descriptors. RFR shows better performances than LASSO for every set; KRR shows slightly better test set errors than RFR, but the RFR predictions on the out-of-sample $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$ points are undeniably better, which gives us confidence to use the random forest models going forward. Figure 4 shows parity plots for the best RFR models trained using 90% of data for training, for ΔH (Cd-rich) in panel a and ΔH (X-rich) in panel b. The corresponding models trained using KRR and LASSO are shown for comparison in Fig. SI-14.

We trained regression models in a similar fashion for ϵ(q/q−1) impurity transition levels. In this case, we add two additional descriptors to the earlier sets: the impurity atom oxidation state (O_1) and the oxidation state (O_2) of the defect site atom (+2 for Cd, −2 for Te/Se/S, and 0 for interstitial), such that $O_1 - O_2 = q$; this enables the training of one model for ϵ(q/q−1), rather than separate models for +2/+1, 0/−1, etc. Fig. 3d,e show the prediction RMSE for test and out-of-sample points, respectively, using the three sets of descriptors (each containing O_1 and O_2 as additional dimensions) as a function of the training set size. While the errors steadily go down with infusion of more training data, there is only a slight improvement in prediction performances going from elemental to unit cell defect properties as descriptors. The respective feature importance values (in %, obtained from the random forest algorithm) have been listed for different RFR models in Table SI-2; it can be seen that while the unit cell defect formation enthalpy has the highest importance for predicting ΔH, as follows from Fig. 3a, the impurity atom oxidation state O_1 shows the highest importance for ϵ(q/q−1). Despite the notable correlation between certain transition levels like +1/0 and 0/−1 and the band edges from unit cell defect calculations, the improvement in prediction performance upon adding unit cell defect properties is less drastic; regardless, the best accuracies are still obtained while using the elemental + unit cell defect properties as descriptors.

Table 2. RMSE (in eV) for regression models trained for PBE ΔH (Cd-rich), using different methods and sets of features.

Dataset	Regression method	Elemental properties	Unit cell defect properties	Elemental + unit cell defect properties
Training	RFR	0.40	0.20	0.17
	KRR	0.40	0.30	0.20
	LASSO	0.62	0.50	0.44
Test	RFR	0.65	0.45	0.38
	KRR	0.68	0.40	0.32
	LASSO	0.75	0.52	0.47
$CdTe_{0.5}Se_{0.5}$	RFR	0.84	0.57	0.52
	KRR	0.80	0.65	0.57
	LASSO	0.95	0.73	0.65
$CdSe_{0.5}S_{0.5}$	RFR	0.86	0.63	0.57
	KRR	0.75	0.68	0.70
	LASSO	0.92	0.70	0.72

From Fig. 3d, e, it can be seen that the RMSE gradually saturates to around 0.31 eV for the test set and 0.34 eV for the out-of-sample points. Further, ϵ(q/q−1) prediction RMSE are listed for RFR, KRR and LASSO models (using 90% of the dataset of CdTe, CdSe, and CdS points for training) in Table 3; KRR predictions when using the elemental + unit cell defect properties are comparably good whereas LASSO errors are higher. Parity plots for the best RFR models trained for ϵ(q/q−1) are presented in Fig. 4, with performances shown (along with the uncertainties) for the training and test points in panel c and for the out-of-sample $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$ points in panel d. Parity plots for models trained using KRR and LASSO are shown in Fig. SI-15.

We have seen that predictive models can be trained for both ΔH and ϵ(q/q−1) using a set of elemental properties and unit cell defect properties as descriptors, and predictions can be made with high accuracy for impurities in out-of-sample mixed-anion compounds. With this confidence, we use the models presented in Fig. 4 to predict the impurity formation enthalpies and charge transition levels (at the PBE level of theory), respectively, for all impurities in CdTe, CdSe, CdS, $CdTe_{0.5}Se_{0.5}$, and $CdSe_{0.5}S_{0.5}$. Before making these predictions for the entire chemical space and using them to screen candidates that act as 'dominating' impurities, we explore the possibility of training such models for the HSE06 $\epsilon(q_1/q_2)$ values. It should be noted that the PBE computed transition levels have been shown to span the physical band gap of the semiconductor[44], and also known to match well with HSE

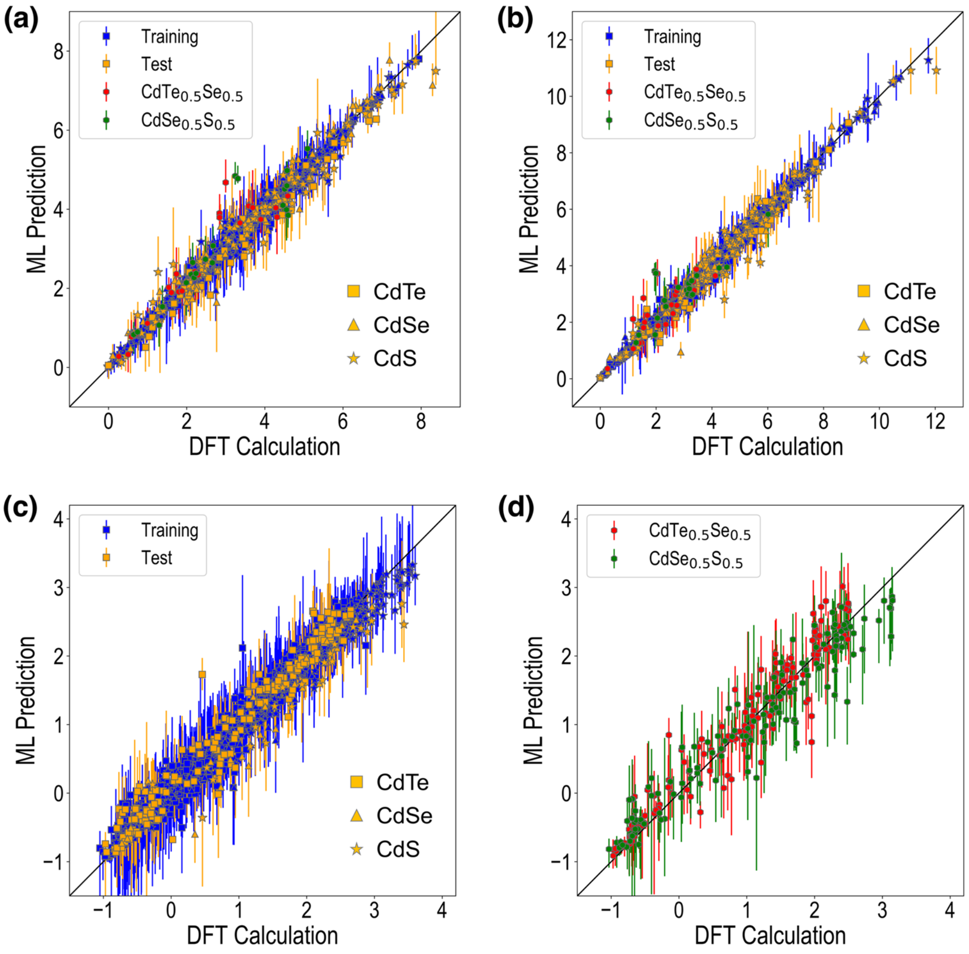

Fig. 4 Trained predictive models. Parity plots for random forest regression models trained for (**a**) ΔH (Cd-rich), and (**b**) ΔH (X-rich). Pictured are the training and test set points (the training set size is 90% of the dataset of CdTe, CdSe, and CdS points), and the $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$ points. Similarly, parity plots are shown for models trained for $\epsilon(q_1/q_2)$ (at the PBE level of theory) using the dataset of 2286 points (total CdTe+CdSe+CdS dataset), for (**c**) the training and test set points, and (**d**) the $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$ points.

Table 3. RMSE (in eV) for regression models trained for PBE $\epsilon(q_1/q_2)$, using different methods and sets of features.

Dataset	Regression method	Elemental properties	Unit cell defect properties	Elemental + unit cell defect properties
Training	RFR	0.18	0.15	0.13
	KRR	0.27	0.28	0.25
	LASSO	0.45	0.42	0.40
Test	RFR	0.34	0.33	0.30
	KRR	0.36	0.35	0.31
	LASSO	0.43	0.40	0.41
$CdTe_{0.5}Se_{0.5}$	RFR	0.35	0.33	0.30
	KRR	0.36	0.30	0.34
	LASSO	0.42	0.34	0.35
$CdSe_{0.5}S_{0.5}$	RFR	0.35	0.34	0.33
	KRR	0.40	0.42	0.37
	LASSO	0.49	0.46	0.44

computed values[54]. As we discuss later, both the PBE and HSE transition levels can compare well with experimentally measured values. In the next section, we present a smaller computational dataset of impurity levels computed using the HSE06 functional, and train predictive models for the same.

DFT data and ML models: HSE $\epsilon(q_1/q_2)$

For the HSE06 impurity calculations, we consider the same chemical space of 63 elements as impurity atoms, and for selected impurities, we calculated 4 transition levels (+2/+1, +1/0, 0/−1, −1/−2), since a large majority of the impurity levels that occur within the band gap or around the band edges belong to one of these 4 transitions. Because of the reliability of PBE formation enthalpies in screening low energy impurities, and the requirement of HSE-based chemical potentials of relevant species, we calculated only $\epsilon(q_1/q_2)$ and not ΔH at the HSE level of theory. As shown in Table 1, we generate computational data for 19% of the total CdTe points, about 10% each of the CdSe and CdS points, and about 5% each of the points belonging to $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$. This totals to a dataset of less than 10% of the entire space of HSE $\epsilon(q_1/q_2)$ levels in the 5 compounds. A glimpse of this dataset is provided in Fig. 5; the +2/+1 to −1/−2 transition levels are plotted for selected impurities in the 5 defect sites in (a) CdTe, (b) CdSe, and (c) CdS. The entire HSE computational data has been plotted in Fig. SI-16 to SI-20.

It is interesting to note from a comparison between Fig. 2 and Fig. 5 that for a given set of impurities, the observed transition

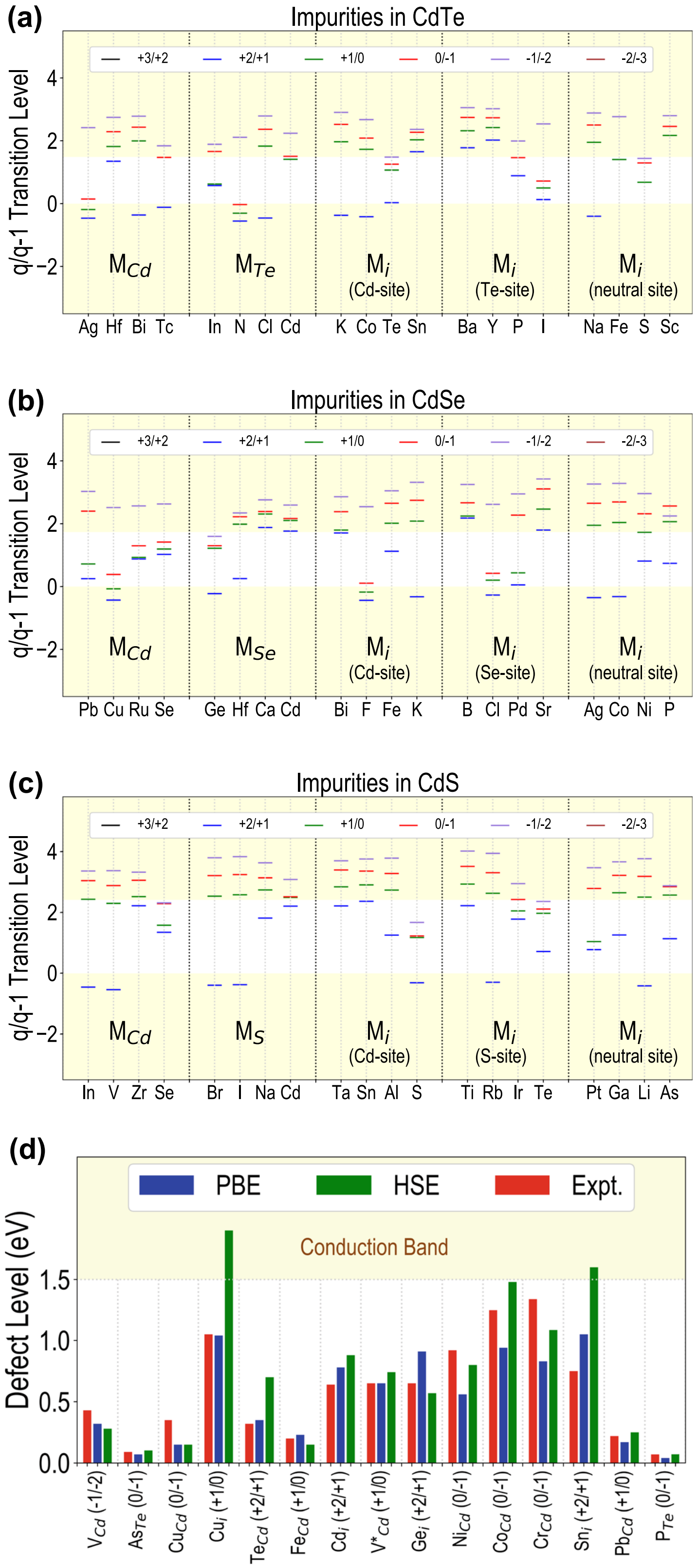

Fig. 5 HSE data and comparison with experiments. Charge transition levels (from +2/+1 to −1/−2) calculated at the HSE06 level of theory for selected impurity atoms in different sites in (**a**) CdTe, (**b**) CdSe, and (**c**) CdS. In **d**, we present a comparison between experimentally measured defect levels[55–60] and the corresponding PBE and HSE computed values in this work. V_{Cd} refers to a Cd vacancy whereas V^*_{Cd} is the Vanadium at Cd site impurity.

levels might occur at different absolute positions but follow the same qualitative trend. For instance, going from Pb_{Cd} to Cu_{Cd} to Ru_{Cd} to Se_{Cd} in CdSe, the +2/+1 impurity level first goes down and then rises towards the CBM in both PBE and HSE. However, Ru_{Cd} and Se_{Cd} exhibit +2/+1 levels deeper in the band gap in HSE than PBE. The same trend can be seen across the PBE and HSE values of the +2/+1 and +1/0 levels for Pt_i, Ga_i, Li_i, and As_i at the neutral interstitial site in CdS. A plot between the PBE and HSE $\epsilon(q_1/q_2)$ in Fig. SI-13 shows that there is a very high correlation between the two; the HSE values lie between the $y = x$ and the $y = x + 1$ lines. We also collected some experimentally measured defect levels in CdTe from the literature[55–60] and plotted a comparison between experiments, PBE $\epsilon(q_1/q_2)$, and HSE $\epsilon(q_1/q_2)$, for various defects in Fig. 5d. It can be seen that in general, there is good correspondence between the three, with the exception of a couple of cases where the HSE value is highly overestimated (Cu and Sn interstitial defects). Based on these 15 data points, PBE $\epsilon(q_1/q_2)$ shows an RMSE of 0.22 eV with respect to experiments, whereas HSE $\epsilon(q_1/q_2)$ shows a higher RMSE of 0.35 eV. There could be many reasons for this discrepancy, such as the requirement of a different mixing parameter[61], but is should be noted that the RMSE for HSE $\epsilon(q_1/q_2)$ drops to 0.18 eV when Cu_i and Sn_i are removed. While the PBE transition levels can be assumed to be reliable, predictions at the HSE level of theory are certainly useful.

We applied the same descriptors as before to train regression models for the smaller dataset of HSE transition levels, but also used the PBE $\epsilon(q_1/q_2)$ as additional descriptors. Similar to Fig. 3a, the linear correlation coefficient plot in Fig. SI-21 shows that while the HSE $\epsilon(q_1/q_2)$ levels have high correlation with certain unit cell defect properties, the correlation between HSE and PBE $\epsilon(q_1/q_2)$ is >0.95. In Fig. 6, we plotted the prediction RMSE as a function of the training set size for the test and out-of-sample sets for RFR models trained for HSE $\epsilon(q_1/q_2)$ using various combinations of descriptors. Figure 6a, b show the errors using the usual three sets of descriptors as before; the performances are nearly identical for the test set across the three descriptor sets, while the unit cell defect properties improve the performances for impurities in $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$. Error saturation is not quite seen when using more than 90% of the CdTe, CdSe, and CdS data for training, which implies that more data is potentially required for training accurate and generalizable models.

In Fig. 6c, d, we plotted the prediction RMSE for the test and out-of-sample points, respectively, using descriptor sets that include the PBE $\epsilon(q_1/q_2)$ values as an added dimension. It can be seen that there is a drastic improvement in prediction performances and both test and out-of-sample errors seem to saturate around 0.24 eV. Further, we trained RFR models for HSE $\epsilon(q_1/q_2)$ using only the PBE $\epsilon(q_1/q_2)$ value as sole descriptor, and see that predictions are similar to the other three sets of descriptors. In Fig. 6e–h, we present four different predictive models; panels e, f, and g show RFR models trained using different sets of descriptors, and it can be seen that the addition of PBE values as descriptors significantly improves the performance. This can also be seen from the RMSE values listed in Table 4–including PBE $\epsilon(q_1/q_2)$ as a descriptor brings down the test and out-of-sample RMSE to ~0.20 eV. We further applied a technique called Delta-learning, wherein we train RFR models for the difference between HSE and PBE transition levels (δ property = HSE $\epsilon(q_1/q_2)$ − PBE $\epsilon(q_1/q_2)$), and predict HSE $\epsilon(q_1/q_2)$ values by adding the predicted δ property to PBE $\epsilon(q_1/q_2)$. It can be seen from Fig. 6(h) that very low test (RMSE = 0.21 eV) and out-of-sample (RMSE = 0.22 eV) errors can be obtained for Delta-learning. Overall, it is seen that the RFR model trained for HSE $\epsilon(q_1/q_2)$ using the elemental properties + unit cell defect properties + PBE $\epsilon(q_1/q_2)$ as descriptors gives the lowest test set and out-of-sample errors, and can be used for making predictions for the >90% of the dataset yet to be computed. Predictive models trained for HSE $\epsilon(q_1/q_2)$ using KRR and LASSO are presented in Fig. SI-22 for comparison with RFR.

Screening of impurities for Fermi level tuning

Using predictions for the neutral state impurity formation enthalpy ΔH, and every impurity transition level $\epsilon(q_1/q_2)$ from +3/+2 to −2/−3, the Fermi level (E_F) and charge (*q*) dependent formation enthalpy (E^f) can be predicted for every possible impurity in Cd-rich or anion-rich chemical potential conditions. For this analysis, we use the machine-learned predictions at the PBE level of theory, since that the formation enthalpies are known to be qualitatively reliable and the transition levels match well with reported experiments, as shown in Fig. 5d. In the absence of any external impurities, the equilibrium Fermi level in a semiconductor is determined by its dominant native point defects, such as vacancies or self-interstitial defects. By comparing the machine-learned formation enthalpy of any impurity with the computed energetics of dominant intrinsic defects, we can estimate the probable change in the nature of conductivity that would occur upon introduction of the impurity in the semiconductor. In order to go through this process, we simulated all possible vacancy (e.g., V_{Cd}, which refers to a Cd vacancy), self-interstitial (e.g., Cd_i or Se_i) and anti-site defects (e.g., Cd_{Te}, S_{Cd}, etc.) in supercells of CdTe, CdSe, CdS, $CdTe_{0.5}Se_{0.5}$, and $CdSe_{0.5}S_{0.5}$. The DFT computed E^f vs. E_F plots for all possible intrinsic defects in the 5 compounds are presented in Figs. SI-23 to SI-27.

The computed energetics of intrinsic defects reveal that while the Cd vacancy, V_{Cd}, is the dominant acceptor type defect in each compound, the Cd interstitial defect, Cd_i(Te-site), is the dominant donor type defect in CdTe, CdSe, and CdS, and Cd interstitial defect, Cd_i(Cd-site), is the dominant donor type defect in $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$. It is also seen that the equilibrium Fermi level (determined using charge neutrality conditions[4]) is near the middle of the band gap for Cd-rich conditions in every compound, which would lead to an intrinsic type of conductivity. The equilibrium E_F shifts towards the valence band upon going from Cd-rich to anion-rich conditions, and in all cases renders the conductivity moderately p-type. If an impurity creates a charged defect that is more stable within the band gap than either the dominant acceptor or donor type defect, it can pin the Fermi level at a different location and change the conductivity. We predicted the E^f values of every possible impurity in the 5 compounds as a function of E_F, and screened those impurities which would cause a shift in the equilibrium E_F. The complete list of all such 'dominating impurities' is provided in Tables SI-1 to SI-5. The dominating defects under Cd-rich and Te-rich conditions and the nature of conductivity are in agreement with reported literature[62].

Given that both DFT and ML predictions of ΔH and $\epsilon(q_1/q_2)$ are available for all 315 possible impurity-site combinations in CdTe, we compare the E^f vs. E_F plots for each impurity estimated from both methods. Specifically, the evaluation of an impurity as shifting or not shifting the equilibrium E_F (alternatively, whether the impurity dominates over the intrinsic defects or not) is used as a metric to compare the DFT and ML predictions. We present such a comparison in Table 5 in terms of the total number of false and true positives or negatives predicted by ML for impurities in Cd-rich and Te-rich chemical potential conditions. It is seen that the false negatives and false positives amount to less than 5% of the total impurities, which means that the ML approach has a >95% probability of successful classification of an impurity as dominating or not. The true positives, which are the impurities predicted to be dominating by both DFT and ML, amount to about 30 in total for both Cd-rich and Te-rich conditions. The total number of dominating impurities as predicted by ML for the 5 compounds (and listed in Tables SI-3 to SI-7) in Cd-rich and anion-rich conditions are presented in Table 6.

For a few selected 'dominating' impurities, we plotted the ML predicted E^f as a function of E_F in Fig. 7 for (a) CdTe, (b) CdSe, (c)

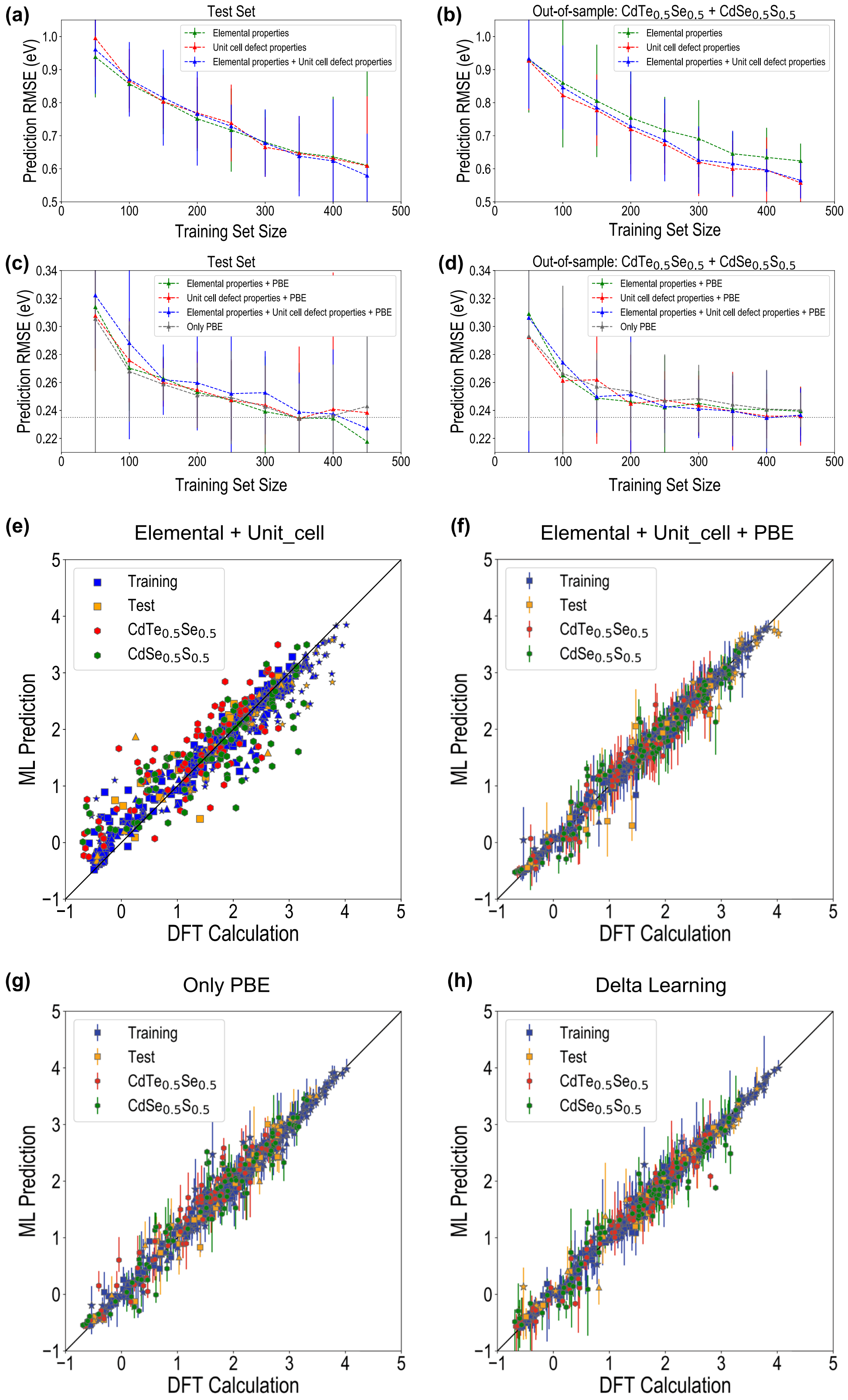

Fig. 6 Predictive models trained on HSE data. Prediction RMSE plotted against the training set size for random forest regression models trained for $\epsilon(q_1/q_2)$ (at the HBE level of theory), using different sets of features, for (**a**) the test set points, without using PBE, (**b**) the out-of-sample points, without using PBE, (**c**) the test set point, using PBE as a descriptor, and (**d**) the out-of-sample points using PBE as a descriptor. Further, parity plots are shown for predictive models trained using 90% of the CdTe+CdSe+CdS dataset as the training set, with performances shown for the training, test and out-of-sample points, using the elemental and unit cell defect descriptors (**e**) without PBE and (**f**) with PBE, (**g**) using just the PBE values as descriptor, and (**h**) using Delta learning. Uncertainties are not plotted in (**e**) because they are very high in general.

Table 4. RMSE (in eV) for regression models trained for HSE $\epsilon(q_1/q_2)$, using different methods and sets of features.

Dataset	Regression method	Elemental properties		Unit cell defect properties		Elemental + unit cell defect properties	
		Without PBE	With PBE	Without PBE	With PBE	Without PBE	With PBE
Training	RFR	0.31	0.10	0.28	0.10	0.28	0.10
	KRR	0.47	0.29	0.40	0.28	0.48	0.28
	LASSO	0.70	0.22	0.63	0.22	0.60	0.21
	δ-learn (RFR)	0.14	0.10	0.13	0.10	0.14	0.09
Test	RFR	0.61	0.23	0.63	0.24	0.62	0.24
	KRR	0.61	0.28	0.57	0.29	0.58	0.30
	LASSO	0.72	0.24	0.65	0.24	0.63	0.24
	δ-learn (RFR)	0.28	0.20	0.28	0.22	0.27	0.21
$CdTe_{0.5}Se_{0.5}$	RFR	0.64	0.22	0.59	0.21	0.58	0.21
	KRR	0.60	0.52	0.54	0.53	0.55	0.53
	LASSO	0.71	0.17	0.54	0.17	0.55	0.16
	δ-learn (RFR)	0.25	0.20	0.20	0.19	0.21	0.18
$CdSe_{0.5}S_{0.5}$	RFR	0.63	0.27	0.61	0.26	0.61	0.26
	KRR	0.57	0.50	0.62	0.51	0.52	0.50
	LASSO	0.71	0.23	0.60	0.22	0.61	0.22
	δ-learn (RFR)	0.28	0.26	0.26	0.25	0.26	0.25

Table 5. A comparison between predictions by DFT and ML of 'dominating impurities' in CdTe.

Verdict	Cd-rich		Te-rich	
	Predicted	% of total	Predicted	% of total
False positives	5	1.59	3	0.95
False negatives	10	3.17	6	1.90
True negatives	272	86.35	275	87.30
True positives	28	8.89	31	9.84

True positives refer to the cases that were predicted to be dominating by both DFT and ML, and true negatives are the cases predicted to be non-dominating by both. False positives were predicted to be dominating by only ML whereas false negatives were predicted to be dominating by only DFT.

Table 6. The total number of impurities predicted to be dominating by ML for Cd-rich and anion-rich chemical potential conditions in the 5 CdX compounds.

CdX	Cd-rich		Te-rich	
	Predicted	% of total	Predicted	% of total
CdTe	28 / 315	8.89	31 / 315	9.84
CdSe	24 / 315	7.62	18 / 315	5.71
CdS	15 / 315	4.76	21 / 315	6.67
$CdTe_{0.5}Se_{0.5}$	44 / 441	9.98	31 / 441	7.03
$CdSe_{0.5}S_{0.5}$	36 / 441	8.16	26 / 441	5.90

CdS, (d) $CdTe_{0.5}Se_{0.5}$, and (e) $CdSe_{0.5}S_{0.5}$, for Cd-rich chemical potential conditions. The formation energies of V_{Cd} and Cd_i are plotted (using dashed lines) as well to illustrate how each impurity dominates and changes the equilibrium E_F. Additional DFT computations (wherever missing) were performed for these selected dominating impurities; the DFT computed E^f is plotted in each case using dotted lines, and it can be seen that there is a very good match between the DFT and ML predicted lines. Impurities such as Na_{Cd}, Zn_{Cd}, F_i, and Cu_{Cd} create acceptor type defects, whereas impurities like Mn_i, Bi_{Cd}, Cl_{Se}, and Li_i are donor type. A common thread across the 5 compounds is low energy defects created by Group I elements and certain transition metals at the Cd-site, halogen atoms and Group V atoms at the X-site, and F, Li, and Ag at the interstitial sites. Indeed, there is abundant experimental literature on using a variety of dopants to change the properties of CdTe, such as p-type doping using As_{Te}[63], Sb_{Te}[64], and Na_{Cd}[65], and improved solar cell efficiency using halogen atoms[66], Zn_{Cd} doping[67], and Li_{Cd} or Li_i[68]. In summary, ML has successfully screened all the impurities that can potentially be introduced in these Cd-chalcogenides to alter the conductivity type and consequently the semiconductor's optoelectronic properties.

Summary

In this work, we showed that machine learning can be used to train accurate predictive models of the formation enthalpy (ΔH) and defect transition levels ($\epsilon(q_1/q_2)$) of impurities in Cd-based chalcogenides using DFT generated data. The choice of descriptors is of vital importance; we see that combining elemental properties of an impurity atom with energetic and electronic information computed from a lower-cost unit cell defect calculation leads to the optimal set of features that serve as inputs to random forest regression models. Predictive models thus trained for ΔH and $\epsilon(q_1/q_2)$ using data generated for CdTe, CdSe, and CdS at the PBE level of theory can accurately predict the impurity properties of mixed anion compounds $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$, demonstrating their out-of-sample predictive power. Models were further trained and tested for a smaller dataset of $\epsilon(q_1/q_2)$ values at the HSE level of theory, for which the use of PBE $\epsilon(q_1/q_2)$ as a descriptor leads to significant improvement in prediction performances. The trained models were used to make predictions for the entire chemical space of impurities in the 5 compounds, following which the formation enthalpy (E^f) of every impurity was obtained as a function of the Fermi level (E_F) in the band gap. The E^f vs. E_F behavior is used to determine whether an impurity can shift the equilibrium E_F in the semiconductor as determined by the dominant intrinsic point defects, leading to a list of impurities

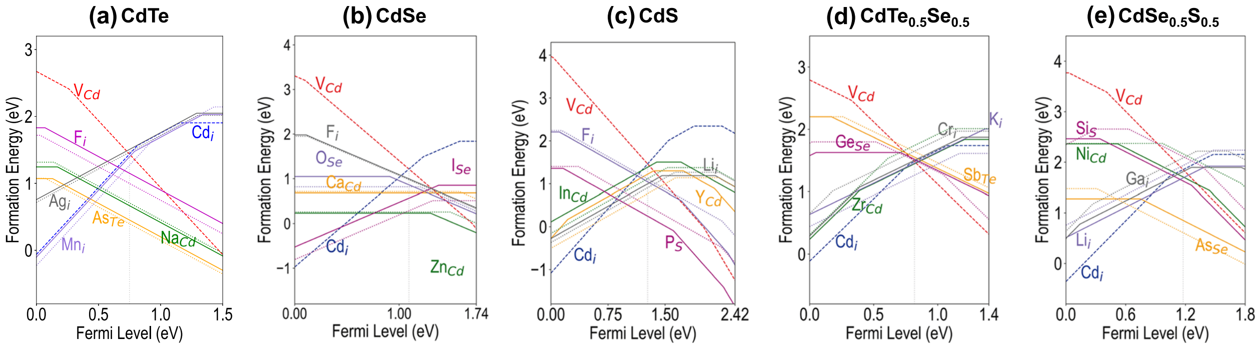

Fig. 7 Predicted impurity energetics. Machine learned defect formation energies at Cd-rich chemical potential conditions for selected impurities predicted to shift the equilibrium Fermi level in (**a**) CdTe, (**b**) CdSe, (**c**) CdS, (**d**) $CdTe_{0.5}Se_{0.5}$, and (**e**) $CdSe_{0.5}S_{0.5}$. The intrinsic defects are shown as dashed lines and ML predictions of different impurities as solid lines, while the dotted lines represent the computed formation energies from DFT; it can be seen that DFT and ML match pretty well.

in each compound that can dominate over the intrinsic defects and change the nature of conductivity in the material. A comparison of DFT and ML predictions shows that less than 5% of the entire population of impurities in CdTe is classified as false negative or false positive (in terms of its 'dominating' nature), giving us confidence that this ML approach can be used for a successful screening of stable and active impurity atoms in preferred defect sites.

The combined DFT and ML approach demonstrated here can be applied to any number of semiconductor classes. For instance, III-V semiconductors such as GaN, GaP, GaAlP, AlP, BP etc. are interesting materials for photodiodes, solar cells, and in recent times, have been studied for intermediate band photovoltaic applications[69–72]. A quick screening of impurity atoms that can not only change the equilibrium Fermi level, but also create energy level(s) in the band gap, can be made possible using machine learned models to predict impurity properties. Given the ubiquity of the descriptors used here, this approach can, in theory, be extended to include all possible pure and mixed compositions of II–VI, III–V, and group IV semiconductors, many of which are currently serving various optoelectronic applications. Further extensions can be made in terms of impurity atoms by including the lanthanides and actinides as well. There are also opportunities in applying a wide variety of descriptors for further improvement in ML performance, such as using Coulomb matrix representation, radial distribution function, or electron density distribution. A true 'semiconductor+impurity' design framework will be complete once the forward prediction model is combined with an inverse model as well, wherein genetic algorithms or other optimization techniques are used to devise suitable compositions which lead to stable impurities with favorable energy levels in the band gap.

METHODS

DFT details

We used 2 × 2 × 2 supercells for any CdX compound, resulting in a system with 64 atoms, to optimize the (fixed cell shape and size) geometry using DFT in the neutral and charged states. The starting structures of CdTe, CdSe, and CdS were obtained from the Materials Project[52]. Anion ordered structures of $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$ were simulated starting from the CdTe and CdSe structures, respectively; the effect of structure on properties was examined by comparing band gaps and selected impurity formation energies for the $CdTe_{0.5}Se_{0.5}$ and $CdSe_{0.5}S_{0.5}$ anion ordered structure, special quasi-random (SQS) structure [49] and a random lower energy structure in Table SI-1 and Fig. SI-30. We find that anion environment has a small (∼0.15 eV or smaller) effect on computed quantities. The computed lattice constants of the 5 compounds are listed in Table SI-1. DFT computations were performed using the Vienna ab-initio Simulation Package (VASP) employing the Perdew-Burke-Ernzerhof (PBE) exchange-correlation functional and projector-augmented wave (PAW) atom potentials. The kinetic energy cut-off for the planewave basis set was 400 eV, and all atoms were relaxed until forces on each were less than 0.05 eV/Å. Brillouin zone integration was performed using a 3 × 3 × 3 Monkhorst-Pack mesh. Further, HSE06 calculations were performed for a smaller dataset using a 4 × 4 × 4 Monkhorst-Pack mesh. The following equations are used to compute the formation enthalpy E^f of an impurity as a function of the chemical potential μ and Fermi level E_F, and any impurity transition level, $\epsilon(q_1/q_2)$:

$$E^f(D^q, E_F) = E(D^q) - E(CdX) + \mu + q(E_F + E_{vbm}) + E_{corr} \qquad (1)$$

$$\epsilon(q_1/q_2) = \frac{E^f(q_1, E_F = 0) - E^f(q_2, E_F = 0)}{q_2 - q_1} \qquad (2)$$

$E(D^q)$ and E(CdX) refer to the total DFT energy of the defect containing system in charge q and the bulk CdX compound, respectively. E_{vbm} refers to the valence band maximum of bulk CdX and E_{corr} is the correction energy necessary due to periodic interaction between charges[14,53].

Regression techniques

RFR is based on ensemble learning through decision trees, where each tree is built using bootstrap samples randomly drawn from the dataset. By optimizing the number of trees and the number of necessary features, RFR prepares a final predictive model as an ensemble, provides errors bars in predictions based on standard deviation across individual trees, and assigns a relative importance to the different features. KRR is a similarity based regression algorithm where the output is expressed as a weighted sum over Kernel functions, which are defined in terms of the Euclidean distance between data points (which is a measure of the similarity). We use a Gaussian kernel in this work, and the hyperparameters that are optimized are the Kernel coefficients and the Gaussian width. LASSO is similar to ridge regression but uses an L1 regularization, unlike KRR which uses L2 regularization. LASSO regression operates on the principle of shrinking the coefficients of many features down to zero, and is thus very useful when there are a large number of features. More details about random forest regression, Kernel ridge regression, and LASSO regression can be obtained from references[32,34,35], respectively. Each technique was applied on the DFT data using the python packages available in Scikit-learn (https://scikit-learn.org/stable/).

DATA AVAILABILITY

DFT data and ML models are available from the corresponding author upon reasonable request.

Received: 2 June 2019; Accepted: 25 February 2020;
Published online: 23 April 2020

REFERENCES

1. Mannodi-Kanakkithodi, A. et al. Comprehensive computational study of partial lead substitution in methylammonium lead bromide. *Chem. Mater.* **31**, 3599–3612 (2019).

2. Shi, T., Yin, W.-J., Hong, F., Zhu, K. & Yan, Y. Unipolar self-doping behavior in perovskite $ch_3nh_3pbbr_3$. *Appl. Phys. Lett.* **106**, 103902 (2015).
3. Park, J. S., Kim, S., Xie, Z. & Walsh, A. Point defect engineering in thin-film solar cells. *Nat. Rev. Mater.* **37**, 194–210 (2018).
4. Sun, R., Chan, M. K. Y., Kang, S. & Ceder, G. Intrinsic stoichiometry and oxygen-induced *p*-type conductivity of pyrite FeS*p*. *Phys. Rev. B* **84**, 035212 (2011).
5. Yan, Y., Yin, W., Shi, T., Meng, W. & Feng, C. *Defect Physics of CH_3 NH_3 PbX_3 (X = I, Br, Cl) Perovskites*, 79–105 (Springer International Publishing, 2016).
6. Luque, A., Martí, A. & Stanley, C. Understanding intermediate-band solar cells. *Nat. Photon.* **6**, 146–152 (2012).
7. Martí, A. et al. Production of photocurrent due to intermediate-to-conduction-band transitions: a demonstration of a key operating principle of the intermediate-band solar cell. *Phys. Rev. Lett.* **97**, 247701 (2006).
8. Heo, S. et al. Deep level trapped defect analysis in $ch_3nh_3pbi_3$ perovskite solar cells by deep level transient spectroscopy. *Energy Environ. Sci.* **10**, 1128–1133 (2017).
9. Rosenberg, J. W., Legodi, M. J., Rakita, Y., Cahen, D. & Diale, M. Laplace current deep level transient spectroscopy measurements of defect states in methylammonium lead bromide single crystals. *J. Appl. Phys.* **122**, 145701 (2017).
10. Robins, L. H., Cook, L. P., Farabaugh, E. N. & Feldman, A. Cathodoluminescence of defects in diamond films and particles grown by hot-filament chemical-vapor deposition. *Phys. Rev. B* **39**, 13367–13377 (1989).
11. Mitsui, T., Yamamoto, N., Tadokoro, T. & Ohta, S. Cathodoluminescence image of defects and luminescence centers in zns/gaas(100). *J. Appl. Phys.* **80**, 6972–6979 (1996).
12. Alkauskas, A., McCluskey, M. D. & Van de Walle, C. G. Tutorial: defects in semiconductors—combining experiment and theory. *J. Appl. Phys.* **119**, 181101 (2016).
13. Brandt, R. E. et al. Searching for "defect-tolerant" photovoltaic materials: combined theoretical and experimental screening. *Chem. Mater.* **29**, 4667–4674 (2017).
14. Freysoldt, C. et al. First-principles calculations for point defects in solids. *Rev. Mod. Phys.* **86**, 253–305 (2014).
15. Herring, C., Johnson, N. M. & Van de Walle, C. G. Energy levels of isolated interstitial hydrogen in silicon. *Phys. Rev. B* **64**, 125209 (2001).
16. Look, D. C., Hemsky, J. W. & Sizelove, J. R. Residual native shallow donor in zno. *Phys. Rev. Lett.* **82**, 2552–2555 (1999).
17. Hofmann, D. et al. Properties of the oxygen vacancy in zno. *Appl. Phys. A* **88**, 147–151 (2007).
18. Suezawa, M. & Sumino, K. Deep defect levels in plastically deformed GaAs. *Jpn. J. Appl. Phys.* **25**, 533–537 (1986).
19. Fernández, P., Piqueras, J., Urbieta, A., Rebane, Y. T. & Shreter, Y. Deformation-induced defect levels in ZnSe crystals. *Semiconductor Sci. Technol.* **14**, 430–434 (1999).
20. Heyd, J., Peralta, J. E., Scuseria, G. E. & Martin, R. L. Energy band gaps and lattice parameters evaluated with the heyd-scuseria-ernzerhof screened hybrid functional. *J. Chem. Phys.* **123**, 174101 (2005).
21. Aryasetiawan, F. & Gunnarsson, O. TheGWmethod. *Rep. Progress Phys.* **61**, 237–312 (1998).
22. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 54 (2017).
23. Xue, D. et al. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, 11241 (2016).
24. Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. Machine learning strategy for accelerated design of polymer dielectrics. *Sci. Rep.* **6**, 20952 (2016).
25. Mannodi-Kanakkithodi, A. et al. Scoping the polymer genome: a roadmap for rational polymer dielectrics design and beyond. *Mater. Today* **21**, 785–796 (2018).
26. Oliynyk, A. O. et al. High-throughput machine-learning-driven synthesis of full-heusler compounds. *Chem. Mater.* **28**, 7324–7331 (2016).
27. Askerka, M. et al. Learning-in-templates enables accelerated discovery and synthesis of new stable double perovskites. *J. Am. Chem. Soc.* **141**, 3682–3690 (2019).
28. Zahrt, A. F. et al. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **363**, 6424 (2019).
29. Tran, K. & Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. *Nat. Catalysis* **1**, 696–703 (2018).
30. Talapatra, A. et al. Autonomous efficient experiment design for materials discovery with bayesian model averaging. *Phys. Rev. Mater.* **2**, 113803 (2018).
31. Kim, C., Pilania, G. & Ramprasad, R. From organized high-throughput data to phenomenological theory using machine learning: the example of dielectric breakdown. *Chem. Mater.* **28**, 1304–1311 (2016).
32. Vu, K. et al. Understanding kernel ridge regression: common behaviors from simple functions to density functionals. *Int. J. Quant. Chem* **115**, 1115–1128 (2015).
33. Vapnik, V. N. *The Nature of Statistical Learning Theory*. (Springer-Verlag, Berlin, Heidelberg, 1995).
34. Couronné, R., Probst, P. & Boulesteix, A.-L. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformat.* **19**, 270 (2018).
35. Gauraha, N. Introduction to the lasso. *Resonance* **23**, 439–464 (2018).
36. Gómez-Bombarelli, R. & Aspuru-Guzik, A. *Machine Learning and Big-Data in Computational Chemistry*, 1–24 (2018).
37. 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).
38. Ferekides, C. & Britt, J. Cdte solar cells with efficiencies over 15%. *Solar Energy Mater. Solar Cells* **35**, 255–262 (1994).
39. Wu, X. High-efficiency polycrystalline cdte thin-film solar cells. *Solar Energy* **77**, 803–814 (2004).
40. Shockley, W. & Queisser, H. J. Detailed balance limit of efficiency of p-n junction solar cells. *J. Appl. Phys.* **32**, 510–519 (1961).
41. Fiducia, T. A. et al. Understanding the role of selenium in defect passivation for highly efficient selenium-alloyed cadmium telluride solar cells. *Nat. Energy* **4**, 504–511 (2019).
42. Wei, S.-H. & Zhang, S. B. Structure stability and carrier localization in CdX (X=S, Se, Te) semiconductors. *Phys. Rev. B* **62**, 6944–6947 (2000).
43. Chan, M. K. Y. & Ceder, G. Efficient band gap prediction for solids. *Phys. Rev. Lett.* **105**, 196403 (2010).
44. Schultz, P. A. Theory of defect levels and the band gap problem in silicon. *Phys. Rev. Lett.* **96**, 246401 (2006).
45. Alkauskas, A., Broqvist, P. & Pasquarello, A. Defect energy levels in density functional calculations: alignment and band gap problem. *Phys. Rev. Lett.* **101**, 046405 (2008).
46. Swanson, D. E., Sites, J. R. & Sampath, W. S. Co-sublimation of cdsexte1x layers for cdte solar cells. *Solar Energy Mater. Solar Cells* **159**, 389–394 (2017).
47. Kim, J.-P., Christians, J. A., Choi, H., Krishnamurthy, S. & Kamat, P. V. Cdses nanowires: compositionally controlled band gap and exciton dynamics. *J. Phys. Chem. Lett.* **5**, 1103–1109 (2014).
48. Yang, J.-H., Yin, W.-J., Park, J.-S., Ma, J. & Wei, S.-H. Review on first-principles study of defect properties of CdTe as a solar cell absorber. *Semiconductor Sci. Technol.* **31**, 083002 (2016).
49. Jiang, Z. et al. Special quasirandom structures for perovskite solid solutions. *J. Phys.* **28**, 475901 (2016).
50. Lingg, M. et al. Structural and electronic properties of cdte1-xsex films and their application in solar cells. *Sci. Technol. Adv. Mater.* **19**, 683–692 (2018).
51. Sankin, I. & Krasikov, D. Kinetic simulations of cu doping in chlorinated cdsete pv absorbers. *Phys. Status Solidi* **26**, 1800887 (2019).
52. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
53. Freysoldt, C., Neugebauer, J. & Van de Walle, C. G. Fully ab initio finite-size corrections for charged-defect supercell calculations. *Phys. Rev. Lett.* **102**, 016402 (2009).
54. Krasikov, D., Knizhnik, A., Potapkin, B., Selezneva, S. & Sommerer, T. First-principles-based analysis of the influence of Cu on CdTe electronic properties. *Thin Solid Films* **535**, 322–325 (2013).
55. Lindström, A., Mirbt, S., Sanyal, B. & Klintenberg, M. High resistivity in undoped CdTe: carrier compensation of te antisites and cd vacancies. *J. Phys. D* **49**, 035101 (2015).
56. Ablekim, T. et al. Self-compensation in arsenic doping of cdte. *Sci. Rep.* **7**, 4563 (2017).
57. Komin, V., Viswanathan, V., Tetali, B., Morel, D.L. & Ferekides, C.S. Investigation of deep levels in cdte/cds solar cells.In *Conference Record of the Twenty-Eighth IEEE Photovoltaic Specialists Conference-2000 (Cat. No.00CH37036)*, 676–679 (2000).
58. Ayoub, M. et al. Annealing effects on defect levels of cdte:cl materials and the uniformity of the electrical properties. *IEEE Trans. Nucl. Sci.* **50**, 229–237 (2003).
59. Jantsch, W. & Hendorfer, G. Characterization of deep levels in cdte by photo-epr and related techniques. *J. Crystal Growth* **101**, 404–413 (1990).
60. Kraft, C. et al. Phosphorus implanted cadmium telluride solar cells. *Thin Solid Films* **519**, 7153–7155 (2011).
61. Pan, J., Metzger, W. K. & Lany, S. Spin-orbit coupling effects on predicting defect properties with hybrid functionals: a case study in cdte. *Phys. Rev. B* **98**, 054108 (2018).
62. Berding, M. A. Native defects in CdTe. *Phys. Rev. B* **60**, 8943–8950 (1999).
63. Ablekim, T. et al. Self-compensation in arsenic doping of CdTe. *Sci. Rep.* **7**, 4563 (2017).
64. Soltani, M., Certier, M., Evrard, R. & Kartheuser, E. Photoluminescence of cdte doped with arsenic and antimony acceptors. *J. Appl. Phys.* **78**, 5626–5632 (1995).
65. Ahmad, F. R. Magnetoresistance in p-type cadmium telluride doped with sodium. *Appl. Phys. Lett.* **106**, 012109 (2015).
66. Dharmadasa, I. M. & Ojo, A. A. Unravelling complex nature of CdS/CdTe based thin film solar cells. *J. Mater. Sci.* **28**, 16598–16617 (2017).

67. Bera, S. R. & Saha, S. Fabrication and characterization of zn-doped cdte nanoparticles based dye sensitized solar cells. *IOSR J. Elect. Electron. Eng.* **11**, 11–18 (2016).

68. Xu, H. et al. Application of lithium chloride dopant in fabrication of cdte solar cells. *J. Electron. Mater.* **46**, 1331–1338 (2017).

69. Zelazna, K. et al. Photoreflectance studies of optical transitions in ganpas intermediate band solar cell absorbers. *Solar Energy Mater. Solar Cells* **188**, 99–104 (2018).

70. Lee, M.-L., Huang, F.-W., Chen, P.-C. & Sheu, J.-K. Gan intermediate band solar cells with mn-doped absorption layer. *Sci. Rep.* **8**, 8641 (2018).

71. BOUMESJED, A. Predicted theoretical efficiency for new intermediate band solar cells (ibsc) based on gaas1-xnx. *J. New Technol. Mater.* **8**, 102–109 (2018).

72. Wahnón, P. & Tablero, C. Ab initio electronic structure calculations for metallic intermediate band formation in photovoltaic materials. *Phys. Rev. B.* **65**, 165115 (2002).

ACKNOWLEDGEMENTS

We acknowledge funding from the US Department of Energy SunShot program under contract DOE DEEE005956. Use of the Center for Nanoscale Materials, an Office of Science user facility, was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. We gratefully acknowledge the computing resources provided on Bebop, a high-performance computing cluster operated by the Laboratory Computing Resource Center at Argonne National Laboratory. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. M.Y.T. would like to acknowledge support from the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program. M.J.D. was was supported by the U. S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences, under Contract No. DE-AC02-06CH11357.

AUTHOR CONTRIBUTIONS

M.K.Y.C., R.F.K., and A.M.K. conceived the idea. A.M.K., M.Y.T., and F.G.S. performed the DFT computations. A.M.K. and M.J.D. trained ML models. All authors contributed to the discussion and writing of the paper.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-0296-7.

Correspondence and requests for materials should be addressed to A.M.-K. or M.K.Y.C.

Reprints and permission 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.