Journal of Materials Chemistry C

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: J. Mater. Chem. C, 2025, 13, 7550

Defect formation in $CsSnI_3$ from density functional theory and machine learning†

Chadawan Khamdang[a] and Mengen Wang *[ab]

Received 10th December 2024,
Accepted 2nd March 2025

DOI: 10.1039/d4tc05215c

rsc.li/materials-c

Sn-based perovskites as low-toxicity materials are actively studied for optoelectronic applications. However, their performance is limited by p-type self-doping, which can be suppressed by substitutional doping on the cation sites. In this study, we combine density functional theory (DFT) calculations with machine learning (ML) to develop a predictive model and identify the key descriptors affecting formation energy and charge transition levels of the substitutional dopants in $CsSnI_3$. Our DFT calculations create a dataset of formation energies and charge transition levels and show that Y, Sc, Al, Zr, Nb, Ba, and Sr are effective dopants that pin the Fermi level higher in the band gap, suppressing the p-type self-doping. We explore ML algorithms and propose training a random forest regression model to predict the defect formation properties. This work shows the predictive capability of combining DFT with machine learning and provides insights into the important features that determine the defect formation energetics.

1 Introduction

Halide perovskites are promising candidates for optoelectronic applications due to their straightforward synthesis methods and optical and charge transport properties.[1–4] The power conversion efficiency (PCE) of Pb-based perovskite-based solar cells (PSCs) has dramatically improved.[5,6] $CsSnI_3$ has been explored as a promising low-toxicity alternative to Pb-based perovskites.[7–10] Despite its potential, the PCE of $CsSnI_3$ remains lower (14.8%)[11] than that of $CsPbI_3$. This reduced efficiency is primarily attributed to the substantial self-p-type doping and defect-assisted nonradiative recombination.[8,12–14]

To address these limitations, defect engineering through doping has been investigated as a potential solution to improve Sn-based perovskite properties. Experimental studies on Ba-doped Sn–Pb perovskites indicate that Ba incorporation can reduce hole concentration, thereby reducing the effects of p-type doping.[15] Density functional theory (DFT) calculations provide a theoretical understanding of the mechanism, showing that Ba acts as an energetically favorable donor in $CsSnI_3$ that shifts the Fermi level upward and decreases the background hole concentration.[16] DFT studies also propose that trivalent cation doping on the Sn site in $MASnI_3$ including Sc, La, and Ce can also raise the Fermi level, which is supported by experimental validation that La doping in $MASnI_3$ results in an increase in photocurrent and open circuit voltage.[17] Another DFT study on $MASnI_3$/$MASnI_2Br$ proposes that Sc, Y, and La doping can shift the Fermi level upward, thereby reducing hole concentration compared to pristine perovskites.[18]

DFT is widely used to predict defect formation energies under various chemical potentials and has reliably predicted intrinsic defect and dopant formation energies and charge transition levels in semiconductors.[13,19–21] Defect calculations require large supercells and hybrid functionals with spin–orbit coupling (SOC) to correctly describe the electronic structure and charge localization, which are computationally demanding.[19,22–24] To overcome these limitations, machine learning (ML) algorithms offer a promising approach to predict and understand defect properties efficiently. Recent studies have demonstrated that DFT can be combined with ML algorithms to predict formation energies and charge transition levels for both dopants and intrinsic defects.[25–28] Specifically for dopant incorporation energetics, data generated from DFT calculations using the PBE functional has been used to train ML algorithms to predict defect formation energies in perovskite oxides (ABO_3) and halide perovskites ($MAPbX_3$).[25,26] There is also a growing interest in applying ML algorithms to predict defect energetics at the hybrid functional accuracy.[27,28] These studies reveal opportunities and the need to improve the prediction of defect formation energetics by combining DFT calculations with hybrid functionals and machine learning methods, which is also promising to provide insights into the physical and chemical descriptors underlying these properties.

This work combines DFT using HSE06 + SOC with ML to predict formation energies and charge transition levels for substitutional dopants in $CsSnI_3$. We explore elements from group II-A (*e.g.*, Mg, Ca), transition metals (*e.g.*, Sc, Y), post-transition metals

[a] Department of Electrical and Computer Engineering, State University of New York at Binghamton, Binghamton, New York 13902, USA.
E-mail: mengenwang@binghamton.edu
[b] Materials Science and Engineering Program, State University of New York at Binghamton, Binghamton, New York 13902, USA

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4tc05215c

(*e.g.*, Al, Ga, In), and metalloids (*e.g.*, Ge, As, Sb). DFT calculations are performed to generate a dataset for formation energies in the neutral (q = 0) and q = +1 charge states as well as the +1/0 charge transition level. We then identify key descriptors affecting formation energy and develop predictive models for the formation energies and charge transition levels of dopants in $CsSnI_3$. Linear and nonlinear regression models including linear regression, Gaussian process regression, kernel ridge regression, and random forest regression are trained. We also analyze the feature correlations and feature importance and extend predictions to other out-of-sample dopants in $CsSnI_3$.

2 Computational details

DFT calculations were performed using the Vienna *Ab initio* Simulation Package (VASP).[29] Projector-augmented wave (PAW) pseudopotentials[30] were employed with a plane-wave energy cutoff of 400 eV. The HSE06 hybrid functional[31] was used with a mixing parameter of 0.54, and the spin–orbit coupling was also included. The Brillouin zone for the primitive unit cell was sampled using a 2 × 2 × 2 Γ-centered k-mesh. The atomic positions were fully relaxed until the forces were less than 0.02 eV $Å^{-1}$. We obtained a lattice constant of a = 8.53 Å, b = 8.81 Å, c = 12.34 Å, and a band gap of 1.32 eV for orthorhombic $CsSnI_3$ with space group *Pnma*, which are in good agreement with experimental values.[8,32] For defect calculations, we used a 2 × 2 × 1 supercell with a 1 × 1 × 2 Γ-centered k-point grid.

The formation energy of a substitutional dopant X on the Sn-site (X_{Sn}) with the charge state of q is calculated by

$$E^f[X_{Sn}^q] = E_{tot}[X_{Sn}^q] - E_{tot}[\text{bulk}] + \mu_{Sn} - \mu_X + q(E_F + E_{vbm}) + E_{corr} \quad (1)$$

$E_{tot}[X_{Sn}^q]$ is the total energy of the supercell containing the substitutional dopant X at charge state q. E_{tot}[bulk] is the total energy of the perfect supercell. E_F is the Fermi level and E_{VBM} is the value for the valence band maximum (VBM). E_{corr} is the Freysoldt's charge correction.[33] μ_{Sn} and μ_X are defined as $\mu_{Sn} = \mu_{Sn}^{bulk} + \Delta\mu_{Sn}$, and $\mu_X = \mu_X^{bulk} + \Delta\mu_X$. μ_{Sn}^{bulk} and μ_X^{bulk} are the single atom energy of the bulk Sn and the dopants. $\Delta\mu_{Sn}$ and $\Delta\mu_X$ are the chemical potentials of Sn and dopants defined by the thermodynamic equilibrium condition of $CsSnI_3$, and against the formation of the competing secondary phases including CsI, SnI_2, SnI_4, and Cs_2SnI_6.

$$\Delta\mu_{Cs} + \Delta\mu_{Sn} + 3\Delta\mu_I = \Delta H_{CsSnI_3} \quad (-5.51\ \text{eV}),$$

$$\Delta\mu_{Cs} + \Delta\mu_I < \Delta H_{CsI} \quad (-3.72\ \text{eV}),$$

$$\Delta\mu_{Sn} + 2\Delta\mu_I < \Delta H_{SnI_2} \quad (-1.65\ \text{eV}),$$

$$\Delta\mu_{Sn} + 4\Delta\mu_I < \Delta H_{SnI_4} \quad (-2.92\ \text{eV}),$$

$$2\Delta\mu_{Cs} + \Delta\mu_{Sn} + 6\Delta\mu_I < \Delta H_{Cs_2SnI_6} \quad (-10.52\ \text{eV}), \quad (2)$$

The numbers in parentheses are the calculated formation enthalpy of the secondary phases using HSE06 + SOC, which show good agreement with the experimental values for $CsSnI_3$ (−5.35 eV), CsI (−3.29 eV), SnI_2 (−1.99 eV), SnI_4 (−2.54 eV), and Cs_2SnI_6 (−9.31 eV).[34] This thermodynamically stable domain of $CsSnI_3$ is illustrated in orange in Fig. 1(b), which is consistent with previous reports.[24] The chemical potentials for I, Sn, and Cs are −0.605 eV, −0.50 eV, and −3.20 eV under I-rich (Sn-poor) condition (point A) and −0.89 eV, 0 eV, and −2.84 eV under I-poor (Sn-rich) condition (point B). We note that $\Delta\mu_X$'s are also determined by the formation of the competing phases XI_n's, where n depends on the oxidation state of the dopant. The data for the formation enthalpy of the XI_n's are made available in the section Data and Code Availability. The charge transition level (CT) from one charged state (q_1) to another (q_2) is defined as

$$CT(q_1/q_2) = \frac{E^f[X_{Sn}^{q_1}, E_F = 0] - E^f[X_{Sn}^{q_2}, E_F = 0]}{q_2 - q_1} \quad (3)$$

Here, $E^f[X_{Sn}^{q_1}, E_F = 0]$ and $E^f[X_{Sn}^{q_2}, E_F = 0]$ are the formation energies calculated at E_F = 0 for the defect in different charge states. The same approach is applied to calculate the formation energy and charge transition level of intrinsic defects in $CsSnI_3$.

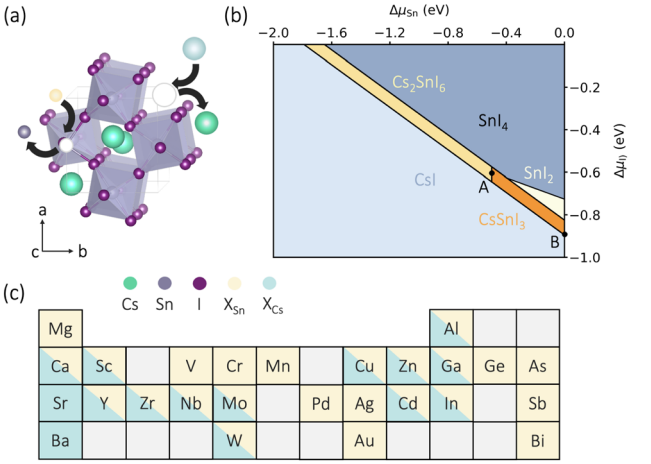

Fig. 1 (a) The structure of orthorhombic $CsSnI_3$ perovskite with the Sn or Cs site substituted by dopants. (b) The thermodynamically stable region for $CsSnI_3$ is shown in orange. Points A and B mark I-rich (Sn-poor) and I-poor (Sn-rich) conditions and the chemical potentials for Sn ($\Delta\mu_{Sn}$) and I ($\Delta\mu_I$). (c) The dopants calculated by DFT.

3 Results and discussions

3.1 Defect formation energy and charge transition level

We performed DFT calculations to obtain the formation energies (E^f) and charge transition levels (CT) for 24 dopants substituting at the Sn site and 15 dopants substituting at the Cs site in $CsSnI_3$, aiming to identify elements that can suppress p-type self-doping. The dopants we calculated are listed in Fig. 1(c), including 4 alkaline earth metals, 15 transition metals, 4 post-transition metals, and 3 metalloids. We calculated the formation energies for all possible charge states, including −1, 0, +1, and +2. The −1 and +2 charge states exhibit higher formation energies and do not appear in the formation energy diagram. As a result, only thermodynamically favorable charge states (q = 0 and q = +1) will

be analyzed. Our search for X_{Sn} is mainly focused on the trivalent dopants, which are expected to be stable at q = +1 under a wide range of the Fermi level, and bivalent dopants, which are expected to be stable at q = 0 under a wide range of the Fermi level. These dopants tend to have shallow or no charge transition levels in the band gap.[17,18,35] To reveal the key features that determine E^f and CT, we also calculated the dopants with different oxidation states, such as Zr, Nb, and Bi.

Fig. 2 includes the intrinsic defects and the dopants with relatively low formation energies under both I-rich [Fig. 2(a)] and I-poor [Fig. 2(b)] conditions. Under the I-rich condition, the E_F determined by the native defects in $CsSnI_3$ is pinned within the valence band (V_B). At VBM, the Cs vacancy (V_{Cs}) at q = −1 has the lowest formation energy, indicating the origin of the p-type self-doping is primarily driven by V_{Cs}, consistent with previous studies.[24] Our DFT study of the $CsSnI_3$ surface phase diagram also shows that surfaces with Cs vacancies are stable under I-rich conditions.[36] Among the calculated dopants, Y_{Sn} at q = +1 has the lowest formation energy at VBM. However, under the I-rich condition, the formation energy of Y_{Sn} at q = +1 is still higher than V_{Cs} at q = −1. Therefore, the Fermi level cannot be shifted to a higher energy under I-rich conditions.

The I-poor condition is preferred to suppress the p-type self-doping. The E_F determined by the native defects is pinned at 0.11 eV above the VBM under the I-poor condition: V_{Cs} with q = −1 is compensated by the I vacancy, which prefers q = +1 near VBM. We note that the low cation vacancy formation energies indicate the low stability of the host material. The cation vacancy formation energies are higher under I-poor and Sn-rich conditions, which benefit the phase stability of $CsSnI_3$. This can be achieved by adding SnF_2 or $SnCl_2$ during the synthesis of the perovskite material, which has been found to slow down the phase transformation to double perovskite and increase the Fermi level.[37] We identified three trivalent elements Al, Sc, and Y that can pin the E_F to higher energies, which are 0.27, 0.32, and 0.33 eV above VBM. Y_{Sn} is only stable at q = +1 in the band gap while Al_{Sn} and Sc_{Sn} have a shallow CT(+1/0) near CBM.

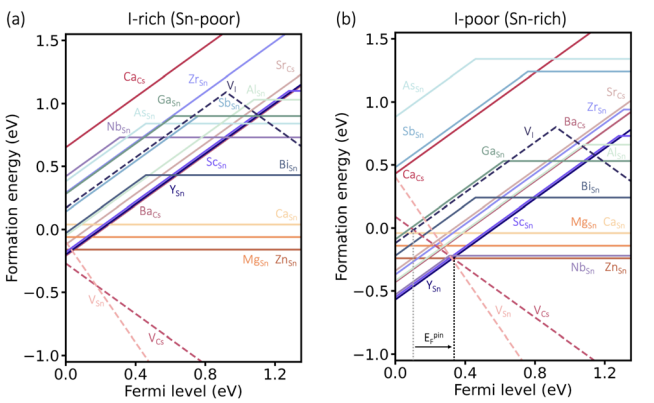

Fig. 2 The calculated defect formation energy diagrams as a function of the Fermi level for native point defects (dotted lines) and substitutional dopants (solid lines) in $CsSnI_3$ under (a) I-rich (Sn-poor) and (b) I-poor (Sn-rich) conditions. The vertical dashed lines (black and grey) indicate the pinned Fermi level (E_F^{pin}).

We confirmed electron localization[20] at the neutral charge state (Fig. S1, ESI†). For example, Fig. S1(a) (ESI†) corresponds to the ground state of Al_{Sn} with the charge localized near the defect while Fig. S1(b) (ESI†) represents a metastable state that is 0.40 eV higher in energy, where the charge is delocalized. When Sn is substituted by bivalent elements including Mg and Zn, the defect is only stable in the neutral charge state and has relatively low formation energies. We also identified two dopants with higher oxidation states (Zr_{Sn} and Nb_{Sn}) that pin the Fermi level above the VBM (~0.2–0.3 eV).

Zr_{Sn} and Nb_{Sn} are stable in the +1 charge state near the VBM, while Nb_{Sn} prefers the neutral charge state across a wide range of the Fermi level. This results in a relatively deeper charge transition level (0.30 eV) within the gap compared to dopants with an oxidation state of 2 or 3. However, we note that deep defects may lead to slow nonradiative recombination rates due to the anharmonicity in perovskite materials.[38] As shown in the density of states (Fig. S2, ESI†), Al_{Sn} and Sc_{Sn} in the q = 0 charge state have localized occupied states near the Fermi level. At q = +1, only delocalized states are observed for the dopants with shallow charge transition levels. Dopants with deep transition levels tend to have a localized unoccupied state in the band gap. For example, Nb_{Sn} and Bi_{Sn} in the q = +1 charge state have an unoccupied state below the CBM, indicating that Nb_{Sn}^+ and Bi_{Sn}^+ can potentially gain an electron to become the neutral charge state.

We now analyze the elemental descriptors of the substitutional dopants that correlate with the target properties including E^f (q = 0), E^f (q = +1), and CT(+1/0) of X_{Sn}, aiming to identify key features to predict these properties. The oxidation state (OS) is an important feature that determines both E^f and CT(+1/0). For elements with OS = 3, the formation energy at q = 0 is higher than the bivalent elements like Zn, Mg, and Ca. The (+1/0) charge transition levels are located near or above CBM for trivalent elements and located below VBM for bivalent elements.

For certain elements with the same OS, there is a direct trend between the atomic radius (AR) of the elements and E^f at both charge states. For example, for Zn, Mg, and Ca with OS = 2, the E^f at q = 0 under the I-rich condition increases [Zn_{Sn} (−0.16 eV) < Mg_{Sn} (−0.06 eV) < Ca_{Sn} (0.04 eV)] while the atomic radius increases from Zn (1.42 Å), Mg (1.45 Å) to Ca (1.94 Å). The trend is consistent for the elements with OS greater than +2. For example, Al has a smaller atomic radius (1.18 Å) than Zr (2.06 Å) and Al has a lower formation energy than Zr in both charge states under I-rich conditions.

The Goldschmidt tolerance factor (t)[39] can be calculated using AR as

$$t = \frac{r_{Cs} + r_I}{\sqrt{2}(r_X + r_I)} \quad (4)$$

where r_{Cs}, r_X, and r_I are the atomic radii of the Cs, X, and I atoms. This factor shows an inverse trend with E^f. Additionally, we find that the octahedral factor ($u = r_X/r_I$)[40] calculated using Shannon's ionic radii (IR)[41] also shows an inverse trend with E^f. Moreover, we also find that the density (D) also shows a direct trend with E^f. The observed trends between AR, D, t, and u and the formation energies are provided in Fig. S3 (ESI†). We note that most of these dopants have a larger AR than Sn.

Electron negativity (EN), ionization energy (IE), and electron affinity (EA) of the dopants play important roles in determining CT(+1/0). For Ca, Mg, and Cu with OS = +2, CT(+1/0) of these dopants are below the VBM following the trend Cu_{Sn} (−0.70 eV) < Mg_{Sn} (−0.30 eV) < Ca_{Sn} (−0.27 eV) and negatively correlated with EN of Cu (0.97) > Mg (0.67) > Ca (0.51). For TMs, CT(+1/0) decreases while the first, second, and third IE increase. For example, CT(+1/0) of Zr_{Sn}, Nb_{Sn}, and Zn_{Sn} are 1.31 eV, 0.30 eV, and −0.27 eV respectively, with 1st, 2nd, and 3rd IE increases from Zr, Nb, to Zn. A similar trend is observed in electron affinity (EA). For instance, the CT(+1/0) levels of Cu_{Sn}, Cr_{Sn}, and Zr_{Sn} are −0.70 eV, −0.29 eV, and 1.31 eV, respectively, with EA decreasing accordingly. The observed correlations of EN, IE, and EA with the charge transition levels are plotted in Fig. S4 (ESI†). The calculated dopants with OS = +4 are not stable in the q = +2 charge state. For example, removing an electron from Zr_{Sn} at q = +1 is energetically unfavorable due to the low energy level of the corresponding occupied state.

In summary, we propose that trivalent dopants including Al, Sc, and Y can raise the Fermi level and suppress the p-type doping of $CsSnI_3$, with Y_{Sn} exhibiting the lowest formation energy under I-poor conditions. Dopants with higher oxidation states, such as Zr and Nb are energetically favorable at q = +1 near the VBM, which also raise E_F to higher values. We also find that formation energies are correlated with properties including the oxidation state, tolerance factor, octahedral factor, and density. Charge transition levels are more correlated with elemental properties including oxidation state, electronegativity, ionization energy, and electron affinity. These observations will guide us in determining features for property predictions using machine learning algorithms.

3.2 Features for machine learning

We initially selected 18 features representing atomic and bulk properties of the substitutional dopants and the corresponding iodide compounds (XI_n). Each feature is expressed as the ratio of the dopant property to the corresponding property of Sn. The atomic and bulk features include the ratios of electronegativity (EN), electron affinity (EA), ionization energy (IE) (including the 1st, 2nd, and 3rd IE), Pauling electronegativity (X), density (D), atomic weight (M), atomic radius (AR), covalent radius (CR), Shannon's ionic radius (IR), and oxidation state (OS) in its most thermodynamically stable substitutional form. We also considered the dopant atomic features including octahedral factor (u), tolerance factor (t), specific heat (S), and heat of vaporization (HV), and thermodynamic properties of XI_n including the heat of formation (HF) from HSE06–SOC calculations [HF(cal)] and experiments [HF(exp)].

We used the Pearson correlation coefficient (p) to identify the features with strong linear correlations with properties and the highly correlated features.[42] If two features have a high absolute Pearson correlation coefficient ($|p| > 0.8$), the one with a low correlation with the property is eliminated from the feature list. In total, 11 features were selected for the ML model training. The correlations between these features and the target properties are shown in Fig. 3. The heatmaps illustrate the relationships between the down-selected features and target properties including E^f (q = 0) [Fig. 3(a)], E^f (q = +1) [Fig. 3(b)], and CT(+1/0) [Fig. 3(c)] under the I-rich condition.

For E^f at q = 0, HV has a positive Pearson correlation coefficient (p = 0.60) with the target property while t(AR) has a negative value of p = −0.31, which is consistent with our observation in Section 3.1. Additionally, the 3rd IE has a strong negative correlation (p = −0.56) and the OS has a strong positive correlation (p = 0.70) with E^f at q = 0. For E^f at q = +1, stronger correlations were observed across most features compared to the other two target properties. Specifically, the HF (exp) exhibited a strong negative correlation (p = −0.55), while D showed a strong positive correlation (p = 0.61). t(AR) has a correlation of −0.20, which is close to the correlation observed in E^f at q = 0 (−0.31). These results indicate that structural stability and physical properties of the dopants are important descriptors to predict E^f.

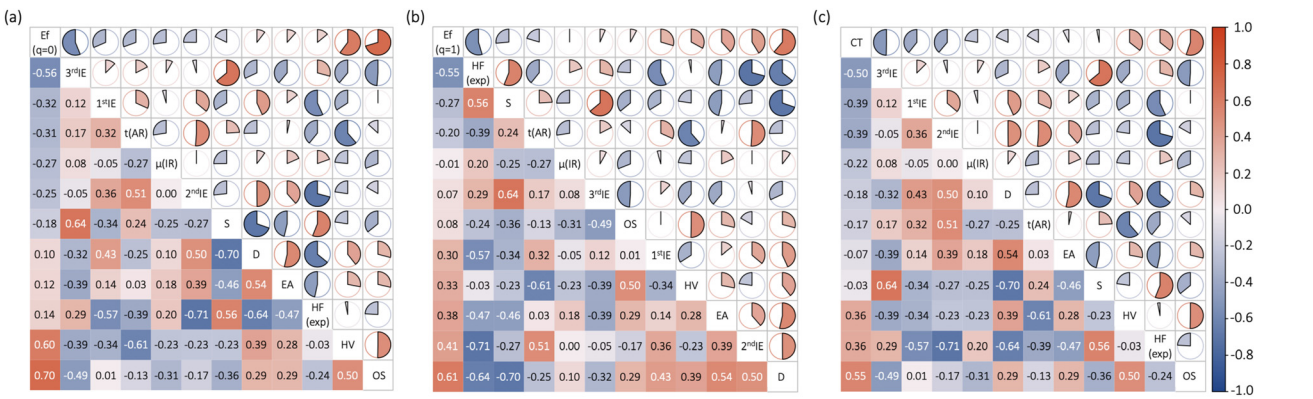

Fig. 3 Pearson correlation matrix capturing pairwise feature–feature and target–feature correlations for the X_{Sn} dataset. The target properties (a) E^f (q = 0), (b) E^f (q = +1), and (c) CT(+1/0) along with the down-selected feature sets are listed on the diagonal. The upper and lower triangular regions of the plot convey the same information in two different visualization schemes. The filled fraction of the pie charts in the upper triangle represents the absolute value of the associated Pearson correlation coefficient, while the lighter and darker shades of color correspond to the strength of the correlation. The matrix of target property-feature correlations ranges from negative to positive correlation, from left to right, or from top to bottom. The features are arranged from strong negative to strong positive correlations, left to right (or top to bottom).

For CT(+1/0), the features with the strongest negative and positive correlations align with those observed in E^f at $q = 0$, as indicated by 3rd IE with $p = -0.50$ and OS with $p = 0.55$ As noted in Section 3.1, the EN is negatively correlated with CT(+1/0) for dopants with an oxidation state of +2. In the selected feature list, EN was excluded due to its high correlation with D (p = 0.79) and the HF (exp) (p = −0.87) and its relatively small variance compared to other elemental properties.

After down-selecting the key features, we trained four machine learning (ML) algorithms including linear regression (LR), Gaussian process regression (GPR), kernel ridge regression (KRR), and random forest regression (RFR) on our DFT dataset to explore their predictive capabilities. In our study, we used the scikit-learn package[43] to train the ML models. We followed standard practices to split the data into training (80%) and testing (20%), apply grid-based hyperparameter search, and employ five-fold cross-validation to reduce overfitting.[25,26] Model performance was evaluated using root mean square error (RMSE) as the key metric. Additionally, we also evaluate feature importance and compare it with the Pearson correlation coefficients.

3.3 Training machine learning models

We first applied the linear regression (LR) model to predict the defect formation energies at $q = 0$ and $q = +1$ and CT(+1/0). The parity plots, training/testing RMSE, and feature importance for the LR model are provided in Fig. S5 (ESI†). The RMSE values for the training/testing data sets were 0.23/0.44 eV for $E^f(q = 0)$, 0.16/0.31 eV for $E^f(q = +1)$, and 0.31/0.45 eV for CT, respectively. Compared with the nonlinear model that will be discussed later, the RMSE of the LR model is higher. Our findings for LR align with previous studies that used linear models to predict defect properties in halide perovskites, where linear regression gives higher RMSEs as compared to nonlinear methods.[26] This highlights the necessity for nonlinear models to fully capture the complexity of defect features and properties.

Gaussian process regression (GPR) is known for modeling complex nonlinear correlations, employing the kernels to define a function based on the covariance of the prior distribution over the target functions.[44,45] We explored five types of kernels and the corresponding hyperparameters, alpha (the regularization parameter), and length to optimize model performance. The kernel functions include Radial Basis Function, ExpSineSquared, Rational Quadratic, DotProduct, and Matern.[46] Hyperparameter optimization was performed using the randomized search method. The optimized hyperparameters are listed in Table S1 (ESI†). The parity plots using GPR are presented in Fig. 4(a), yielding training/testing RMSE values of 0.21/0.32 eV for $E^f(q = 0)$, 0.18/0.23 eV for $E^f(q = +1)$, and 0.16/0.31 eV for CT.

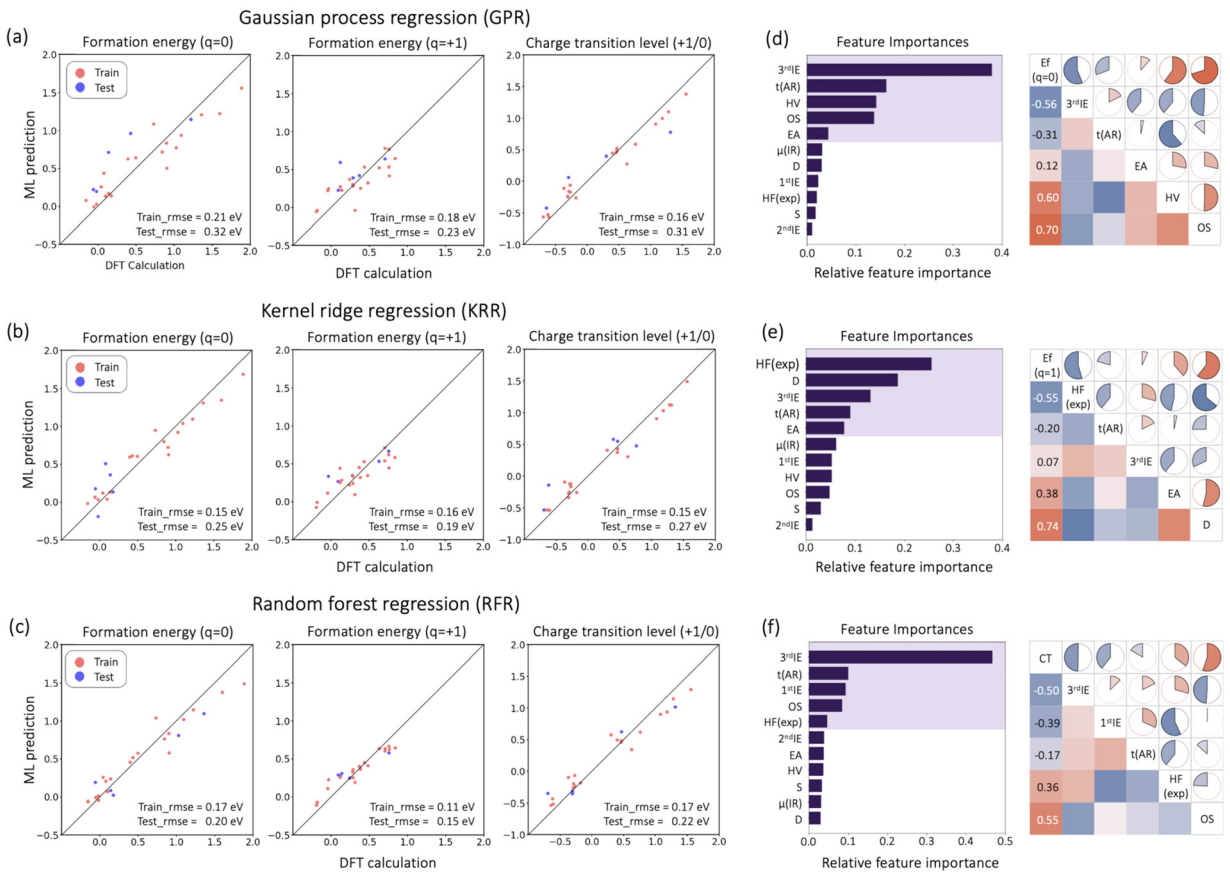

Fig. 4 Parity plots from (a) Gaussian process regression, (b) kernel ridge regression, and (c) Random forest regression. The relative feature importance from random forest regression for predicting (d) formation energy ($q = 0$), (e) formation energy ($q = +1$), and (f) charge transition level [CT(+1/0)]. The shaded purple region highlights the top five most important features.

GPR outperformed LR for all three target properties, indicating its effectiveness in capturing the underlying relationships in the data.

Kernel ridge regression (KRR) is also a nonlinear regression model integrating ridge regression with kernel functions.[47] The same kernel functions were tested as in GPR. The best estimators for KRR result in RMSE values of 0.15/0.25 eV for $E^f(q = 0)$, 0.16/0.19 eV for $E^f(q = +1)$, and 0.15/0.27 eV for CT, as shown in Fig. 4(b).

Random forest regression (RFR) is a widely used machine learning technique that combines multiple decision trees into an ensemble of predictors.[48] Training the RFR model involves optimizing hyperparameters including the number of trees (or estimators), maximum tree depth, number of leaf nodes, and the maximum number of features used to split a tree. The best hyperparameters that yielded the best predictions for all regressions are listed in Table S1 (ESI†). The parity plots from the RFR model are shown in Fig. 4(c). The RMSE for the training/testing datasets are 0.17/0.20 eV for $E^f(q = 0)$, 0.11/0.15 eV for $E^f(q = +1)$, and 0.17/0.22 eV for CT, respectively. These results demonstrate improved predictions for both E^f and CT(+1/0) compared to those achieved by using LR, GPR, and KRR.

During the training of the RFR model, we also assessed the feature importance for the three target properties [Fig. 4(d)–(f)]. For $E^f(q = 0)$ [Fig. 4(d)], the top five most important features from the RFR training are 3rd IE, t(AR), HV, OS, and EA. For $E^f(q = +1)$ [Fig. 4(e)], the top five most important features include HF(exp), D, 3rd IE, t(AR), and EA. The feature importance for predicting formation energy for both charge states highlights three important features: 3rd IE, t(AR), and EA. These features exhibit relatively strong positive or negative Pearson correlations in Fig. 3 and partially overlap with the top important features predicting E^f of neutral defects in ABO_3.[25] For CT(+1/0) [Fig. 4(f)], the top import features include 3rd IE, t(AR), 1st IE, OS, and HF(exp). These features are also consistent with the highly correlated features shown in Fig. 3(c).

3.4 Prediction with random forest regression

We also trained RFR and KRR using the formation energies of X_{Sn} under the I-poor condition, aiming to directly predict out-of-sample dopants that can suppress the p-type self-doping. The RMSE values for the training/testing datasets of RFR are 0.22/0.28 eV for $E^f(q = 0)$ and 0.16/0.21 eV for $E^f(q = +1)$, which are lower than KRR as shown in Fig. S6 (ESI†). The top important features for $q = 0$ [3rd IE, HV, and OS] and $q = +1$ [HF(exp), D, and 3rd IE] remain consistent with the top important features derived from the formation energies calculated under the I-rich condition [Fig. 4(d) and (e)]. We apply the trained RFR model to predict the formation energies of 23 out-of-sample dopants under the I-poor condition. The formation energies for $q = 0$ and $q = +1$ are provided in Table S2 (ESI†). Our predictions indicate that there are three trivalent dopants (La, Ce, and Pr) with the $E^f(q = +1)$ lower than that of V_I at $q = +1$ and the CT(+1/0) level is close to CBM. This suggests that these dopants can shift the Fermi level, pinning it closer to the conduction band compared to the intrinsic Fermi level at 0.11 eV above VBM. Additionally, Sr and Ba with OS = 2 have relatively low formation energies at $q = 0$, which is consistent with previous DFT calculations using the HSE06 functional, confirming the predictive capability of the RFR model for formation energy.[16] We performed DFT calculations for La_{Sn} and Ce_{Sn}, as shown in Fig. S7 (ESI†), which found that the E^f of La_{Sn} is lower than that of Ce_{Sn}. This is consistent with the ML predictions. The prediction of the formation energy at $q = +1$ performs better than $q = 0$. These calculations are in good agreement with previous calculations using the HSE06 functional, which show that La and Ce doping in $MASnI_3$ raises the Fermi level due to the low formation energy at $q = +1$ and no charge transition level in the band gap.[17]

3.5 Substitutional dopants on both Sn and Cs sites from random forest regression

We performed DFT calculations of 15 substitutional dopants on the Cs site. Ba_{Cs} and Sr_{Cs} are only stable at $q = +1$ within the band gap. Ba_{Cs} has the lowest formation energy and will pin the Fermi level at 0.26 eV under the I-poor conditions, which is also consistent with the previous study on alkaline-earth metal doping at the Cs site.[16] We also find Sr_{Cs} to have low formation energy, pinning the Fermi level at 0.21 eV under the I-poor conditions.

We also applied RFR to predict the formation energy at $q = +1$ under I-rich conditions for substitutional dopants on both the Cs site (X_{Cs}) and the Sn site (X_{Sn}). Fig. 5(a) shows the parity plot of DFT calculated *versus* the RFR predicted values for E^f ($q = +1$) with a train/test RMSE of 0.17/0.26 eV. This RMSE is higher than that of the X_{Sn} system as the features need to describe the interaction between dopants with two cation sites. The top five features from the RFR training are shown in Fig. 5(b), including HV, OS, D, EA, and 1st IE. Two of these features [D and EA] are consistent with the top features from RFR training using only the X_{Sn} data points [Fig. 4(e)], indicating the consistency in feature correlations on both sites.

4 Conclusion

In conclusion, we performed DFT calculations using the HSE06 functional with SOC to identify substitutional dopants in $CsSnI_3$ that suppress the p-type self-doping. Trivalent dopants

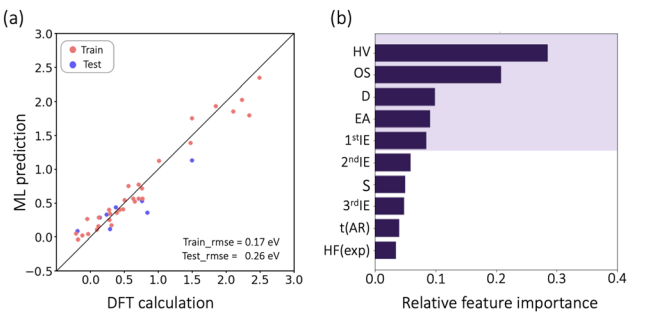

Fig. 5 (a) Parity plot obtained from random forest regression and (b) highlights the top five most important features for the formation energy ($q = +1$) of X_{Cs} and X_{Sn}.

including Al_{Sn}, Sc_{Sn}, and Y_{Sn} prefer the +1 charge state and have shallow or no charge transition levels in the band gap, which pin the Fermi level at 0.27, 0.32, and 0.33 eV under the I-poor conditions. Bivalent dopants including Mg and Zn are only stable in the neutral charge state and have low formation energies. We also identified the dopants with a high oxidation state, Zr_{Sn} and Nb_{Sn}, which can also raise the Fermi level under the I-poor condition. For the substitutional dopants on the Cs site, we identified Ba_{Cs} and Sr_{Cs} that are only stable in the q = +1 charge state and can pin the Fermi level at 0.26 and 0.21 eV under the I-poor condition.

We explore machine learning regression algorithms and determine that the random forest regression can be used to develop a predictive model for the formation energy and charge transition levels of substitutional defects at the cation sites in $CsSnI_3$. By analyzing the feature correlation and feature importance from the random forest regression training, we identified key features including oxidation state, the heat of formation, density, and ionization energy as key descriptors that determine the defect formation energetics. The trained model is also applied to predict out-of-sample dopants and predicts three dopants including La, Ce, and Pr that have low formation energies at the q = +1 charge state. From a theoretical perspective, this study identifies key features that predict formation energy and charge transition levels. We believe that this predictive model will be valuable for investigating defects that suppress p-type behavior in other Sn-based perovskite materials, and provide insights into the key elemental descriptors that determine the energetics in defect formation.

Data availability

Datasets containing the defect formation energies and chemical potentials and the ML codes for training and prediction are available from https://github.com/Mengen-W/Doping_CsSnI3_DFT_ML.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The work was supported by the new faculty start-up and Transdisciplinary Areas of Excellence (TAE) Seed Grant funds from SUNY Binghamton. This work used Bridges-2 at Pittsburgh Supercomputing Center through allocation MAT230043 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. This work also used computational resources provided by the SPIEDIE cluster at the State University of New York at Binghamton.

References

1 Y. Zhou and Y. Zhao, *Energy Environ. Sci.*, 2019, **12**, 1495–1511.
2 Q. Dong, Y. Fang, Y. Shao, P. Mulligan, J. Qiu, L. Cao and J. Huang, *Science*, 2015, **347**, 967–970.
3 S. D. Stranks, G. E. Eperon, G. Grancini, C. Menelaou, M. J. Alcocer, T. Leijtens, L. M. Herz, A. Petrozza and H. J. Snaith, *Science*, 2013, **342**, 341–344.
4 J. H. Noh, S. H. Im, J. H. Heo, T. N. Mandal and S. I. Seok, *Nano Lett.*, 2013, **13**, 1764–1769.
5 A. Kojima, K. Teshima, Y. Shirai and T. Miyasaka, *J. Am. Chem. Soc.*, 2009, **131**, 6050–6051.
6 National Renewable Energy Laboratory (NREL). Best research-cell efficiency chart. (2022)., https://www.nrel.gov/pv/cell-efficiency.html.
7 I. Chung, B. Lee, J. He, R. P. Chang and M. G. Kanatzidis, *Nature*, 2012, **485**, 486–489.
8 I. Chung, J.-H. Song, J. Im, J. Androulakis, C. D. Malliakas, H. Li, A. J. Freeman, J. T. Kenney and M. G. Kanatzidis, *J. Am. Chem. Soc.*, 2012, **134**, 8579–8587.
9 G. E. Eperon, S. N. Habisreutinger, T. Leijtens, B. J. Bruijnaers, J. J. van Franeker, D. W. DeQuilettes, S. Pathak, R. J. Sutton, G. Grancini and D. S. Ginger, *et al.*, *ACS Nano*, 2015, **9**, 9380–9393.
10 A. Babayigit, A. Ethirajan, M. Muller and B. Conings, *Nat. Mater.*, 2016, **15**, 247–251.
11 B.-B. Yu, Z. Chen, Y. Zhu, Y. Wang, B. Han, G. Chen, X. Zhang, Z. Du and Z. He, *Adv. Mater.*, 2021, **33**, 2102055.
12 T.-B. Song, T. Yokoyama, C. C. Stoumpos, J. Logsdon, D. H. Cao, M. R. Wasielewski, S. Aramaki and M. G. Kanatzidis, *J. Am. Chem. Soc.*, 2017, **139**, 836–842.
13 P. Xu, S. Chen, H.-J. Xiang, X.-G. Gong and S.-H. Wei, *Chem. Mater.*, 2014, **26**, 6068–6072.
14 T. Shi, H.-S. Zhang, W. Meng, Q. Teng, M. Liu, X. Yang, Y. Yan, H.-L. Yip and Y.-J. Zhao, *J. Mater. Chem. A*, 2017, **5**, 15124–15129.
15 Z. Yu, X. Chen, S. P. Harvey, Z. Ni, B. Chen, S. Chen, C. Yao, X. Xiao, S. Xu and G. Yang, *et al.*, *Adv. Mater.*, 2022, **34**, 2110351.
16 J. Zhang and L. Chen, *J. Phys. Chem. Lett.*, 2023, **14**, 4058–4062.
17 L. Gregori, C. Frasca, D. Meggiolaro, P. Belanzoni, M. W. Ashraf, A. Musiienko, A. Abate and F. De Angelis, *ACS Energy Lett.*, 2024, **9**, 3036–3041.
18 L. Gregori, D. Meggiolaro and F. De Angelis, *Small*, 2024, 2403413.
19 J. Kang, J. Li and S.-H. Wei, *Appl. Phys. Rev.*, 2021, **8**, 031302.
20 S. Mu, M. Wang, J. B. Varley, J. L. Lyons, D. Wickramaratne and C. G. Van de Walle, *Phys. Rev. B*, 2022, **105**, 155201.
21 C. Freysoldt, B. Grabowski, T. Hickel, J. Neugebauer, G. Kresse, A. Janotti and C. G. Van de Walle, *Rev. Mod. Phys.*, 2014, **86**, 253–305.
22 X. Zhang, M. E. Turiansky, J.-X. Shen and C. G. Van de Walle, *Phys. Rev. B*, 2020, **101**, 140101.
23 X. Zhang, M. E. Turiansky, J.-X. Shen and C. G. Van de Walle, *J. Appl. Phys.*, 2022, **131**, 090901.
24 J. Zhang and Y. Zhong, *Angew. Chem., Int. Ed.*, 2022, **134**, e202212002.
25 V. Sharma, P. Kumar, P. Dev and G. Pilania, *J. Appl. Phys.*, 2020, **128**, 034902.

26 A. Mannodi-Kanakkithodi and M. K. Chan, *J. Mater. Sci.*, 2022, **57**, 10736–10754.
27 A. Mannodi-Kanakkithodi, X. Xiang, L. Jacoby, R. Biegaj, S. T. Dunham, D. R. Gamelin and M. K. Chan, *Patterns*, 2022, **3**, 100450.
28 J. B. Varley, A. Samanta and V. Lordi, *J. Phys. Chem. Lett.*, 2017, **8**, 5059–5063.
29 G. Kresse and J. Furthmüller, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 11169.
30 P. E. Blöchl, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1994, **50**, 17953.
31 J. Heyd, G. E. Scuseria and M. Ernzerhof, *J. Chem. Phys.*, 2003, **118**, 8207–8215.
32 Z. Chen, C. Yu, K. Shum, J. J. Wang, W. Pfenninger, N. Vockic, J. Midgley and J. T. Kenney, *J. Lumin.*, 2012, **132**, 345–349.
33 C. Freysoldt, J. Neugebauer and C. G. Van de Walle, *Phys. Rev. Lett.*, 2009, **102**, 016402.
34 J. E. Saal, S. Kirklin, M. Aykol, B. Meredig and C. Wolverton, *JOM*, 2013, **65**, 1501–1509.
35 M. A. Irham, F. H. Tejo Baskoro, F. A. Permatasari and F. Iskandar, *J. Phys. Chem. C*, 2022, **126**, 5256–5264.
36 K. Li, C. Khamdang and M. Wang, *Phys. Rev. Mater.*, 2024, **8**, 093401.
37 S. Gupta, D. Cahen and G. Hodes, *J. Phys. Chem. C*, 2018, **122**, 13926–13936.
38 J. Zhang, X. Zhang, M. E. Turiansky and C. G. Van de Walle, *PRX Energy*, 2023, **2**, 013008.
39 V. M. Goldschmidt, *Naturwissenschaften*, 1926, **14**, 477–485.
40 C. Li, X. Lu, W. Ding, L. Feng, Y. Gao and Z. Guo, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2008, **64**, 702–707.
41 R. D. Shannon, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 1976, **32**, 751–767.
42 I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang and I. Cohen, *Noise Reduction in Speech Processing*, 2009, pp. 1–4.
43 F. Pedregosa, *J. Mach. Learn. Res.*, 2011, **12**, 2825.
44 M. Seeger, *Int. J. Neural Syst.*, 2004, **14**, 69–106.
45 J. L. Puga, M. Krzywinski and N. Altman, *Nat. Methods*, 2015, **12**, 277–278.
46 D. Duvenaud, PhD thesis, University of Cambridge, 2014.
47 V. Vovk, *Empirical inference: Festschrift in honor of vladimir n. vapnik*, Springer, 2013, pp. 105–116.
48 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.