THE JOURNAL OF PHYSICAL CHEMISTRY C
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCC

Article

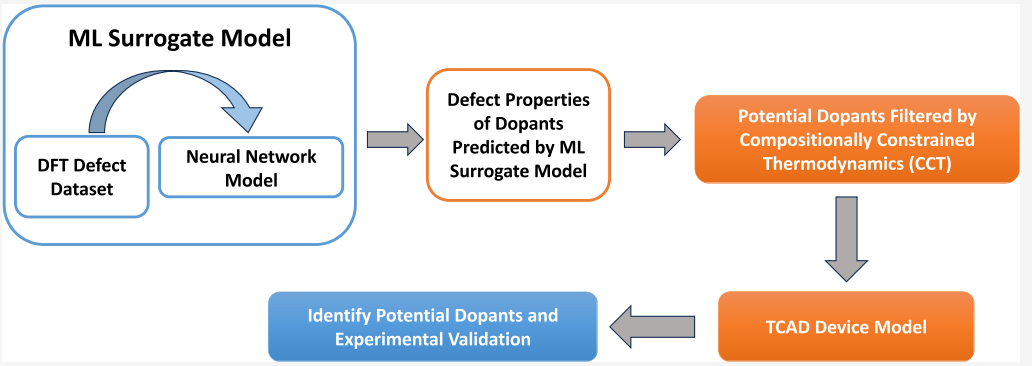
Exploration of Deep Learning Models for Accelerated Defect Property Predictions and Device Design of Cubic Semiconductor Crystals

Xiaofeng Xiang,* Dylan Soh, and Scott Dunham*

Cite This: *J. Phys. Chem. C* 2024, 128, 8821−8829

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: In this work, we present an exploration of deep learning models for predicting defect properties in cubic phase semiconductors. The nature of impurity energy levels strongly influences the performance of semiconductors in a wide range of applications, such as solar cells, field effect transistors, and qubits for quantum computing. In this work, we employ two types of deep learning models, a crystal defect graph neural network and a chemical environment-encoded artificial neural network, to predict defect properties. The models are trained on a data set of charge-dependent defect formation energies obtained from density functional theory computations and descriptors based on elemental properties, defect local environment, and relevant semiconductor properties. We assess the models' performance and showcase their capability in optimizing semiconductor devices, particularly when used in tandem with compositionally constrained thermodynamics and technology computer-aided design models.

INTRODUCTION

Point defects play a crucial role in semiconductor devices as they can significantly affect device performance by facilitating atomistic diffusion[1] or creating trapping centers.[2] These defects can be categorized into two types on the basis of their transition levels relative to the valence band maximum (VBM) and conduction band minimum (CBM): shallow defects and deep level defects. Low concentrations of shallow level defects generally have minimal impact on device performance, whereas extrinsic concentrations of shallow defects are desirable dopants as they can be effectively ionized to control p-type or n-type doping. On the other hand, deep level defects can introduce trapping centers, thereby limiting carrier lifetime through Shockley−Read−Hall (SRH) recombination. Consequently, a comprehensive understanding of the electronic properties of defects is critical for semiconductor device research. Density functional theory (DFT) calculations have emerged as a valuable tool for characterizing defect properties and uncovering the doping mechanisms in semiconductor devices.[3] Nevertheless, due to the necessity of extensive structural relaxation and the possible presence of many different elements, the computational cost associated with studying defects can be extremely high.

In recent years, machine learning (ML) methods have emerged as promising tools for rapidly predicting material properties with comparable accuracy to DFT. Substantial progress has been made in developing ML models for property predictions in both molecules and crystals. Among these models, graph neural networks (GNNs) have shown superior performance due to their ability to capture structural information from crystals or molecules, implicitly incorporating many−body interactions.[4−7] This approach has also been applied to predict defect properties in metal oxides and 2D materials.[8,9] However, when it comes to the prediction of defect properties in bulk materials, the performance of GNN

Received: February 21, 2024
Revised: May 3, 2024
Accepted: May 8, 2024
Published: May 16, 2024

models is still inferior compared to models specifically designed for bulk materials. This discrepancy hinders their wider application in this domain. This is primarily due to the substantial deviation of relaxed defect structures from the ideal crystal lattice, posing challenges in accurate predictions.[10,11]

In our previous work,[12] we introduced a ML framework for predicting defect formation energies and charge transition levels in a diverse range of cubic phase semiconductors. We evaluated the performance of various ML models and identified the best-performing model, which enabled efficient screening of deep level defects potentially induced by extrinsic dopants. Our screening accuracy exceeded 95%; however, the limitations of the ML framework prevent quantitative studies due to errors that are much larger than the thermal energy at relevant temperatures. This results in carrier density prediction errors exceeding 10^5 in some cases, rendering it unsuitable for an accurate high-throughput prediction model. For rigorous quantitative analysis, we must constrain the error to less than 1 order of magnitude, necessitating a prediction error within around 0.2 eV.[13]

In this current work, we aim to enhance our model using deep learning techniques. First, we incorporate the chemical environment of defects using the smooth overlap of atomic positions (SOAP) representation[14] into our existing artificial neural network (ANN) model. We compare this model with a fine-tuned crystal graph convolutional neural network (CGCNN) model[4] for defect predictions. Our findings reveal that both models significantly improve the prediction accuracy of formation energies by around 50%. Among them, the chemical environment-encoded ANN model outperforms both formation energy and transition level predictions. Specifically, the prediction errors for II−IV semiconductors are all lower than 0.21 eV, making it suitable for quantitative high-throughput screening.

Moreover, we seamlessly integrate our deep learning model with CCT[1] and TCAD technology[15] to investigate potential dopants for enhancing semiconductor device performance. This collaborative approach harnesses the predictive capabilities of our model and TCAD to identify optimal dopants, thereby optimizing device functionality and enhancing overall performance. To validate the efficacy of this design framework, we conduct tests on CdTe solar cell devices and compare to experimental data. Within this framework, we predict four dominant figures of merit: efficiency (η), open-circuit voltage (V_{oc}), short-circuit current (J_{sc}), and fill factor (FF) for solar cells with various potential dopants in the chemical space. These predictions enable us to determine the most promising dopants that have the potential to significantly impact solar cell performance.

METHODS

Training Data Set. Our data set encompassed 1640 doped semiconductors, incorporating AB-type compounds where "A" denotes the cation and "B" signifies the anion, across groups II−VI, III−V, and IV−IV. This categorization yielded 8 II−VI compounds (such as CdO, CdS, CdSe, CdTe, ZnO, ZnS, ZnSe, and ZnTe), 16 III−V compounds (including BN, BP, BAs, BSb, AlN, AlP, AlAs, AlSb, GaN, GaP, GaAs, GaSb, InN, InP, InAs, and InSb), and 10 group IV compounds (such as C, Si, Ge, and Sn and binary combinations like SiC, GeC, SnC, SiGe, SiSn, and GeSn). These compounds, totaling 34, were modeled adopting the cubic zinc blende structure, characterized by A atoms forming a face-centered cubic (FCC) lattice and B atoms occupying tetrahedral sites.

Within any given AB compound crystallized in the zinc blende structure, potential defects could emerge at the A-site, B-site, or multiple, symmetrically distinct interstitial positions. The present investigation examines five defect locations: A-site, B-site, A-site tetrahedral interstitial (surrounded by four A atoms), B-site tetrahedral interstitial (surrounded by four B atoms), and the neutral site hexagonal interstitial (equidistant from three A and three B atoms). For the binary compounds, all the five defect sites are considered, whereas for the four elemental systems (C, Si, Ge, and Sn), three defect sites are analyzed (A-site, A-site interstitial, and neutral site interstitial).

In terms of the doping elements, we take into account a wide spectrum from periods I to VI, along with all lanthanides, culminating in 77 unique species. As a result, the overall count of potential impurities within this chemical structure is 12,474. Of these, around 10% have been calculated via DFT to obtain neutral state formation energies under both A-rich and B-rich conditions and six charge transition levels [$\epsilon(-1/0)$, $\epsilon(-2/-1)$, $\epsilon(-3/-2)$, $\epsilon(+1/0)$, $\epsilon(+2/+1)$, and $\epsilon(+3/+2)$]. Following the application of PCA for the removal of outliers, our training data set includes formation energies for 1476 compounds and transition levels across 1076 configurations. The data for this study are obtained from our previous research.[12]

Throughout the research process, the physical and chemical descriptors gathered and encoded from our previous work[12] were added to the training data set to be used by the neural networks. Such descriptors were based on intrinsic properties from a variety of atomic impurities across the periodic table; inherent properties of cubic-structured binary IV−IV, III−V, and II−VI semiconductors; and the Coulomb matrix which mimics the electrostatic interaction around defect sites.

Furthermore, we collected six inner averages of SOAP power spectrum descriptors[16] with a single radial basis function, degree 1 spherical harmonics, and a cutoff distance of 6 Å. These descriptors accurately describe the invariant atomic positions within each semiconductor system by considering the permutations of gathering atomic pairs from the atomic species that compose the defect supercell structure. The input defect structures for ML models are pristine semiconductor supercells with introduced defects, but without any structural relaxation. We experimented with increasing the number of radial basis functions and degrees of spherical harmonics in SOAP, but we observed minimal improvement in neural network models, prompting us to stick with the mentioned configuration.

Chemical Environment-Encoded ANN. We constructed our chemical environment-encoded ANN once all categorical and continuous features from the training data set were normalized and applied with one-hot encoding. The algorithm was trained on the descriptors mentioned previously to learn and optimize the features to predict defect formation energies and transition levels. Each NN architecture contains two to three dense neuron layers, through which the input is concatenated before returning the output through the final layer.

To determine the regression network's effectiveness in predicting each of the eight outputs (two types of formation energies and six different transition levels), we utilized a 5-fold stratified K-fold cross-validation sampling on the population of semiconductors in the input data set based on the type of semiconductor and defect atom type. Such sampling and validation ensure that the predictive results from the network

https://doi.org/10.1021/acs.jpcc.4c01124
J. Phys. Chem. C 2024, 128, 8821−8829

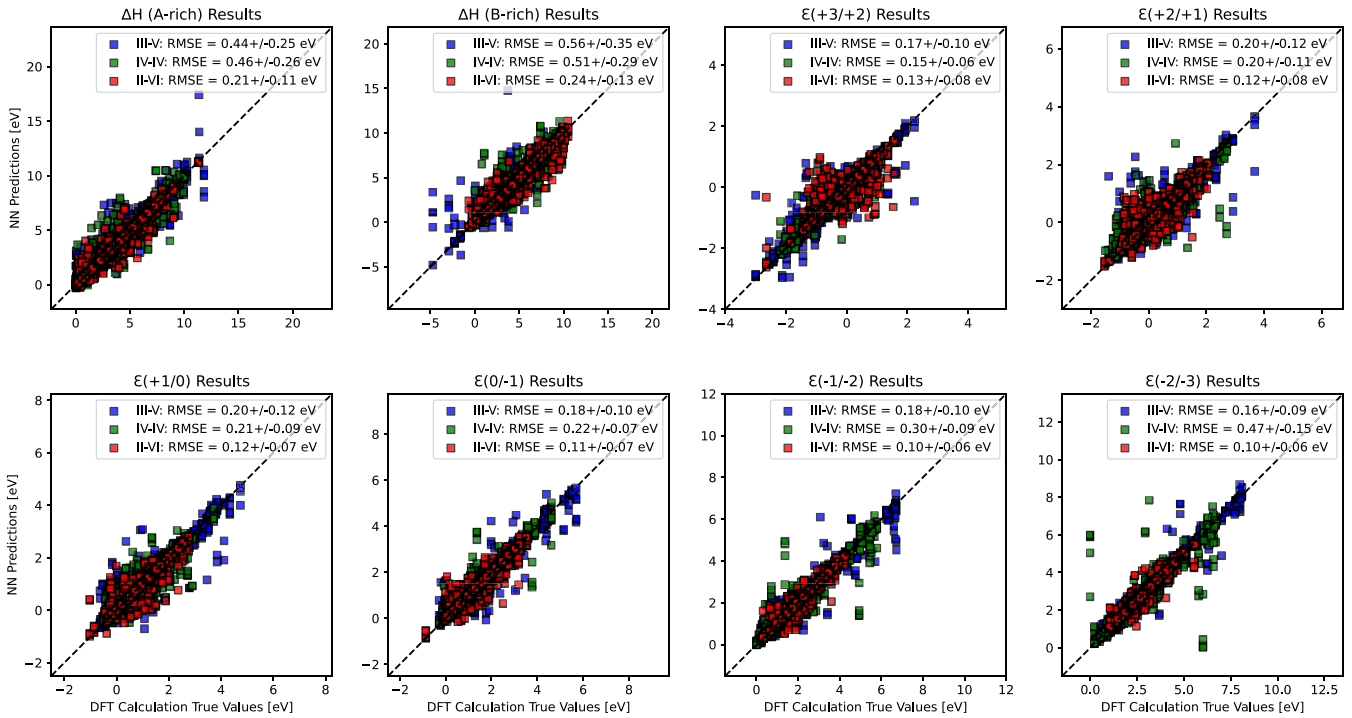


Figure 1. Parity plots for the ANN.

are limited in bias and variance, reducing overfitting. Hyperparameter optimization was performed via Bayesian optimization to minimize the cross-validation error for the average of each of the five splits; such parameters include learning rate, learning rate decay for the Adam optimizer, the standard deviation for the Gaussian noise distribution, and the kernel regularizer.

From each of the five splits from the stratified *K*-fold cross-validation, our baseline measurement used and calculated to determine the network's performance was the root-mean-square error (RMSE) compared to the predictions and actual DFT values. All training and testing RMSEs for each fold were then averaged. This procedure was repeated for another five runs or trials, in which each run's uncertainties were averaged together.

This procedure gives us a practical testing error obtained for every input data point, providing us with the accurate predictive of the network on each of the eight output values. The optimal set of hyperparameters is chosen to minimize the cross-validation error; we ultimately report training and test errors for every model, but optimization is based on the validation error, such that the actual test set in each iteration remains unseen by the model during the training process.

To better emphasize our testing RMSEs in principle, we crafted parity plots (shown in Figure 1) that compare the benchmark DFT-computed properties to the predicted values from the neural network. As such, we present and visualize the testing results on the A-rich and B-rich formation energies and the six charge transition levels, labeled with the reported average testing RMSEs in eV from Table 1 and their standard deviations, for each binary semiconductor type. For each semiconductor group, labeled with different colors for emphasis, all testing data points were cataloged from every individual five stratified *K*-fold method splits across all the five runs.

Table 1. ANN Model Performance

predictors	training error (eV)	testing error (eV)	II−VI error (eV)	III−V error (eV)	IV−IV error (eV)
ΔH (A-rich)	0.032	0.31	0.21	0.44	0.46
ΔH (B-rich)	0.036	0.36	0.24	0.56	0.51
$\epsilon(-1/0)$	0.018	0.16	0.11	0.18	0.22
$\epsilon(-2/-1)$	0.024	0.18	0.10	0.18	0.30
$\epsilon(-3/-2)$	0.042	0.24	0.10	0.16	0.47
$\epsilon(+1/0)$	0.012	0.16	0.12	0.20	0.21
$\epsilon(+2/+1)$	0.010	0.15	0.12	0.20	0.20
$\epsilon(+3/+2)$	0.010	0.14	0.13	0.17	0.15

It can be observed that the integration of descriptors and SOAP power spectrum features significantly enhanced the predictive performance. This improvement is quantified through the average RMSEs obtained from five independent runs, which are systematically tabulated for various output categories as delineated in Table 1. Notably, the RMSEs for formation energies exhibited over 60% enhancement, while the improvements for charge transition levels, specifically $\epsilon(-1/0)$, $\epsilon(-2/-1)$, $\epsilon(+1/0)$, $\epsilon(+2/+1)$, and $\epsilon(+3/+2)$, approached nearly 50% relative to the precedent model benchmarks presented in Table SII. The performance metrics for $\epsilon(-3/-2)$ exhibited only marginal advancements. These outcomes substantiate the significant impact of the SOAP features in refining the model's accuracy.

In the "training error" column of Table 1, we recorded low RMSE values for all output properties across various semiconductor groups, suggesting that the model effectively assimilates the descriptor data from the input and reliably predicts these outputs. Nevertheless, a notable discrepancy between training and testing predictions raises the concern of potential overfitting. To assess the robustness of our models,

https://doi.org/10.1021/acs.jpcc.4c01124
J. Phys. Chem. C 2024, 128, 8821−8829

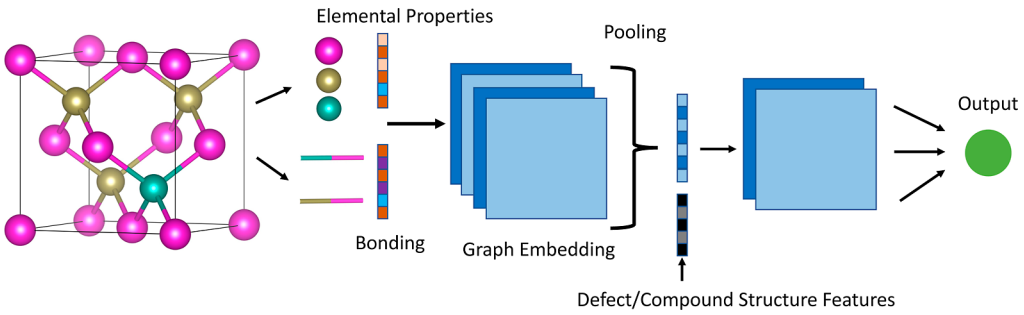


Figure 2. Schematic of CDGNN.

we determined the standard deviation through a 5-fold cross-validation process. As depicted in Figure 1, the standard deviation hovers at approximately half the magnitude of the prediction error, a variance deemed acceptable by prevailing reports in the literature.[9,17] Although a larger regularization term might bridge the test-train performance gap, such a modification was deemed unnecessary. The regularization parameter has been finely tuned via Bayesian optimization, and further increases might introduce undue bias. Moreover, the model's predictions align closely with experimental data in subsequent evaluations of dopant efficacy in devices, reinforcing our confidence in its exceptional performance.

In terms of the network's performance on different outputs, it displays a minor improvement in predicting the formation energies of doped semiconductor systems that are A-rich compared to the B-rich ones. In addition, it is also shown that group II−VI binary group semiconductors have the lowest RMSEs and have most of their data points in line with the diagonal of the parity plot for both formation energy types compared with the III−V and IV−IV groups.

For each binary semiconductor type in each column of Table 1, the RMSE values relatively change regardless of the charge transition level type being measured (with the notable exception of $\epsilon(-1/-2)$ and $\epsilon(-2/-3)$ for IV−IV binary group semiconductors). For instance, the RMSEs for II−VI semiconductors in each individual transition level remain roughly between 0.10 and 0.13 eV, while for III−V semiconductors, the errors remain between 0.16 and 0.20 eV. Additionally, it is noted that for the binary group II−VI semiconductors, the network again predicts their transition levels effectively compared to their counterparts, as observed from the low error values and having the majority of their data points remain close to the diagonal, which show reduced variance from the actual DFT values.

Crystal Defect Graph Neural Network. Our CDGNN follows closely with the original crystal graph neural network (CGCNN) framework of ref 4. We create a surrogate model, F_{CDGNN}, parametrized by weight $\mathbf{W}$, which has the form

$$\hat{y}_{\mathrm{d}} = F_{\mathrm{CDGNN}}(G_{\mathrm{d}}, \mathbf{u}_{\mathrm{elem}}, \mathbf{u}_{\mathrm{struc}}; \mathbf{W}) \tag{1}$$

where G_{d} refers to the graph of unrelaxed defect structures, $\mathbf{u}_{\mathrm{elem}}$ denotes the elemental properties of atoms comprising the defect structures, and $\mathbf{u}_{\mathrm{struc}}$ refers to compound properties (e.g., band gap, compound formation enthalpy, and compound lattice constant) as well as defect structure properties (e.g., Coulomb matrix, SOAP descriptors). The function F_{CDGNN}, parametrized by weights $\mathbf{W}$, effectively maps a defected crystal G_{d} along with its corresponding features to the target property $\hat{y}_{\mathrm{d}}$. In this study, $\hat{y}_{\mathrm{d}}$ are defect formation energy and transition levels.

As depicted in Figure 2, the defect elemental properties and bonding information are initially embedded in the feature vector $\mathbf{v}_i$. The convolutional layers then iteratively update $\mathbf{v}_i$ by incorporating information from surrounding atoms and bonds, employing a nonlinear graph convolution function (eq 2). After T convolutions, the CDGNN learns a comprehensive feature vector $\mathbf{v}_{\mathrm{d}}$ that represents the unrelaxed defect crystal structure through pooling all nodes in the structure (eq 3).

Next, the CDGNN incorporates the compound and defect structure features, such as supercell lattice parameter, bandgap, and SOAP features, by concatenating them with the graph-embedded nodes (eq 4). As shown in Table 2, we observe an improvement of around 10% in the RMSEs after encoding the structure features. This improvement is achieved through several hidden layers, leading to the final output. Upon implementing the graph of defects and their corresponding features into the model, we observed a significant improvement in RMSEs of formation energy predictions compared to our previous models in Table SII, achieving an RMSE of approximately 0.5 eV (Table 2) using the same train-test splitting method from the ANN model. However, it is worth noting that no substantial improvement was observed in the transition level predictions. This may be due to the CDGNN model including a large number of features, such as bonding and element properties of all atoms, which requires a sufficient amount of data for training. To improve the CDGNN model's performance, future work will focus on either incorporating a feature pruning method or increasing the amount of training data

Table 2. CDGNN Model Performance

predictors	training error (eV)	testing error (eV)
ΔH (A-rich) w/o SOAP	0.103	0.589
ΔH (A-rich)	0.117	0.537
$\epsilon(-1/0)$	0.176	0.463
$\epsilon(-2/-1)$	0.420	0.488
$\epsilon(-3/-2)$	0.299	0.391
$\epsilon(+1/0)$	0.148	0.391
$\epsilon(+2/+1)$	0.169	0.406
$\epsilon(+3/+2)$	0.146	0.350

$$\mathbf{v}_i^{t+1} = \mathrm{conv}(\mathbf{v}_i^t, \mathbf{v}_j^t, \mathbf{u}_{(i,j)_q}), \qquad \mathbf{u}_{(i,j)_q} \in G_{\mathrm{d}} \tag{2}$$

$$\mathbf{v}_{\mathrm{d}} = \mathrm{Pool}(\mathbf{v}_0^{(T)}, \mathbf{v}_1^{(T)}, \ldots\ldots, \mathbf{v}_N^{(T)}) \tag{3}$$

$$\mathbf{v}_{\mathrm{d}}' = \mathbf{v}_{\mathrm{d}} \oplus \mathbf{u}_{\mathrm{struc}} \tag{4}$$

https://doi.org/10.1021/acs.jpcc.4c01124
J. Phys. Chem. C 2024, 128, 8821−8829

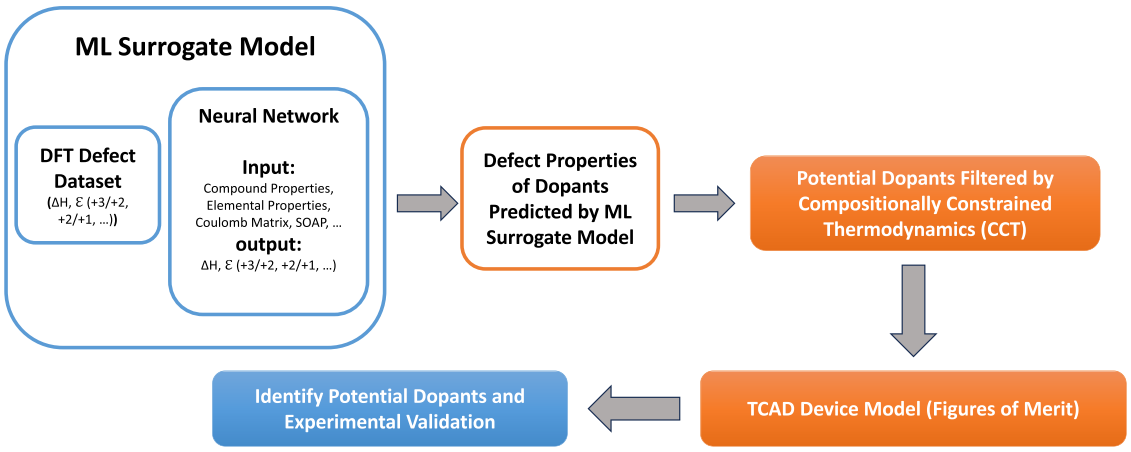


Figure 3. Schematic of ML-assisted device design framework. DFT is employed to build a comprehensive defect property database including defect formation energy and thermodynamic transition levels. A surrogate neural network model is trained using this data set to predict defect properties for dopants across the entire chemical space. Next, CCT is used to filter and select dopants with desired properties. Then, the defect information on these potential dopants is incorporated into the TCAD model of the specific device of interest, allowing us to gather figures of merit. In the end, we identify promising potential dopants that demonstrate potential for enhancing device performance. Experimental testing and validation are conducted to assess the suitability and efficacy of the selected dopants.

ML-Assisted Device Design Framework. Most state-of-the-art ML models for defects focus on predicting the electronic properties of individual defects. However, in complex real-world systems, such as transistors and solar cells, multiple defects associated with a dopant coexist and collectively influence device performance. Consequently, a comprehensive evaluation framework for dopants is essential to effectively screen and identify optimal dopants. In response to this need, we propose a novel ML-assisted device design framework (Figure 3) that facilitates quantitative and efficient screening and exploration of dopants in semiconductor devices. This innovative framework streamlines the process of material discovery and device design, accelerating the exploration of new dopants for advanced semiconductor devices.

In this framework, the ML-assisted device design framework operates as follows: first, we employ DFT to construct a comprehensive defect property database, encompassing defect formation energy and thermodynamic transition levels. Subsequently, this data set is used to train a surrogate neural network model capable of predicting defect properties for dopants throughout the entire chemical space. Next, we apply compositionally constrained thermodynamics (CCT)[18] to filter and select dopants with desired properties. The defect information from these potential dopants is then incorporated into the TCAD model specific to the device of interest, facilitating the assessment of essential figures of merit. By leveraging this framework, we are able to identify promising potential dopants with the capacity to enhance device performance. To validate the predictions, experimental testing and validation are needed to ensure the suitability and efficacy of the selected dopants for practical device applications.

CCT is a canonical approach to calculate the defect concentrations. Under a dilute approximation, defect concentrations are given by

$$x_{k,q} = \frac{\theta_{k,q}}{N^{\text{total}}} e^{-\Delta H^{f}_{k,q}/k_B T} \tag{5}$$

where $\theta_{k,q}$ is the degeneracy factor that counts the number of equivalent defect configurations, $\Delta H^{f}_{k,q}$ is the formation energy of defect k at charge q, and N^{total} is the total number of lattice sites in a perfect material. One can obtain deviations of atomic fractions f_α from perfect stoichiometry f^p_α under given defect concentrations in the material using the following relation

$$f_\alpha = \frac{f^p_\alpha - \sum_{k,q} x_{k,q} n^\alpha_k}{1 - \sum_{k,q} x_{k,q} \sigma_{k,q}} \tag{6}$$

where n^α_k is the number of atoms of type α added or removed from the system when one defect k exists. $\sigma_{k,q} = \sum_\alpha n^\alpha_k$ is the atom amount differences in a system with and without the defect k. If atomic fractions and standard reference chemical potentials are provided, one can determine the defect concentration, chemical potentials for each element, and the Fermi level by solving the set of equations defined by (6). We refer to this method as CCT.[18]

In this study, the atomic fractions of the semiconductor compound and its impurities are unknown. Instead, we can use the chemical potentials of each element in the semiconductor compound and its potential dopants to determine the defect concentration and Fermi level. If the formation energies $H^{(r)}_{k,q}$ are known for given reference values $\mu^{(r)}_\alpha$ and $E^{(r)}_F$, then $H_{k,q}$ for any values of μ_α and E_F can be expressed using the following equations

$$\Delta H_{k,q}(\mu_\alpha, E_F) = \Delta H^{(r)}_{k,q} + \sum_\alpha n^\alpha_i(\mu_\alpha - \mu^{(r)}_\alpha) + q(E_F - E^{(r)}_F) \tag{7}$$

Using eqs 5 and 7, we can represent the total defect density for any defect using the density of a single neutral defect

$$x_k = x_{k,0} \sum_q \frac{\theta_{k,q}}{\theta_{k,0}} e^{\frac{-\Delta H^{(r)}_{k,q} - \Delta H^{(r)}_{k,0} + q(E_F - E^{(r)}_F)}{k_B T}} \tag{8}$$

From 5 and 8, we can easily obtain

$$\sum_\alpha n^\alpha_k \mu_\alpha = \sum_\alpha n^\alpha_k \mu^{(r)}_\alpha - k_B T \ln \frac{x_{k,0} \times N^{\text{total}}}{\theta_{k,0}} - \Delta H^{(r)}_{k,0} \tag{9}$$

https://doi.org/10.1021/acs.jpcc.4c01124
J. Phys. Chem. C 2024, 128, 8821−8829

Table 3. ML-Assisted Dopant Screening for Device Design (Example of CdTe Solar Cells)

dopant	condition	dominant defect	deep level?	type	Fermi level (eV)	carrier density @ 300 K (cm^{-3})	carrier density in experiment (cm^{-3})	device performance			
								η (%)	V_{oc} (V)	J_{sc} (mA)	FF
	Te-rich	V_{Cd}, Te_{Cd}	yes	p-type	0.34	7.53×10^{13}	10^{13} to 10^{15}[10,34,35]	16.59	0.86	22.56	85.69
F	Te-rich	F_int_A	no	p-type	0.12	1.98×10^{17}	10^{14} to 10^{17}[32]	20.94	1.07	22.31	88.06
N	Cd-rich	N_Te	no	p-type	0.25	1.16×10^{15}	10^{15} to 10^{17}[36]	18.45	0.85	22.56	87.42
P	Te-rich	P_Te	no	p-type	0.20	9.70×10^{15}	10^{15} to 10^{17}[34,37]	19.55	0.90	22.53	87.61
As	Te-rich	As_Te	no	p-type	0.24	1.61×10^{15}	10^{15} to 10^{17}[37−39]	18.54	0.94	22.56	87.17
As	Cd-rich	As_Te	no	p-type	0.23	2.22×10^{15}	10^{15} to 10^{17}[37−39]	18.55	0.95	22.59	87.44
Cu	Te-rich	Cu_Cd	no	p-type	0.14	8.37×10^{16}	10^{14} to 10^{16}[29,40,41]	20.53	1.05	22.29	88.10

By applying eq 9 and the charge neutrality eq 10, we can determine the equilibrium defect distribution and the Fermi level for any given set of chemical potentials and temperature.

$$n + \sum_k \sum_{q<0} |q| x_{k,q} = p + \sum_k \sum_{q>0} q x_{k,q} \tag{10}$$

Taking CdTe as an example, a commonly used semiconductor material in thin-film solar cells,[19] we employ the above approach to screen dopants' potential impact on device performance. First, we utilize defect formation energies calculated via DFT for intrinsic defects and employ the ANN model to predict formation energies and transition levels for extrinsic defects. With this information, we determine the equilibrium defect concentrations for different charge states at a given temperature via CCT. Considering the typical manufacturing process of CdTe solar cells, which involves high-temperature vapor deposition followed by cooling and annealing at room temperature,[20] we select a temperature of 893 K for the equilibrium defect concentration calculations, followed by a quench to room temperature. We assume a "frozen-in" approximation,[21] where the total defect concentrations are held constant from the prior equilibrium calculation. Defects are then redistributed among available charge states based on the new Fermi level and temperature.

Once we obtain the defect distribution, we incorporate these defects into a well-constructed CdTe TCAD device model. Here, defects are categorized as either shallow or deep levels. For deep levels (V_{Cd} and Te_{Cd}[22−26]), we take Shockley−Read−Hall (SRH) recombination into account, while for extrinsic defects, we exclude SRH recombination due to low capture rates. The capture rates of deep levels are computed using the NONRAD method.[27,28] Further details about the device model are provided in the Supporting Information (Table SI). In the end, we generate a dopant screening table, as shown in Tables 3 and SIII. For each dopant, we explore two conditions in CdTe: Cd-rich and Te-rich. And the chemical potential for each dopant is the maximum limit of their lowest formation energy binary or ternary compounds formed with Cd or Te. This provides an optimistic or pessimistic estimation based on the dopants' solubility in CdTe. The table includes key characteristics of the defects, such as dominant defects, whether they introduce deep levels (with energy levels within the band gap ±0.2 eV), the material type after doping, Fermi level relative to the VBM at 300 K, and the majority carrier density predicted by the ML framework (negative values indicating electron density and positive values indicating hole density). Experimental reports of majority carrier density at 300 K are also included for comparison. Additionally, the TCAD model provides four key performance metrics for solar cell devices: efficiency (η), open-circuit voltage (V_{oc}), short-circuit current (J_{sc}), and fill factor (FF). These metrics offer a clear indication of the dopants' potential to enhance device performance.

Using these features, we efficiently screen potential dopants to enhance device performance. Table 3 presents a selection of promising dopants for CdTe solar cells, which do not introduce deep-level defects and significantly increase carrier density compared to intrinsic CdTe, suggesting that the formation energies of dominant extrinsic defects of these dopants are more favorable compared to the intrinsic defects under specified conditions. In the first row, intrinsic CdTe is included as a reference. We list several dopants which are able to raise carrier density several orders compared to intrinsic CdTe. Remarkably, this framework exhibits excellent performance, selecting some of the most commonly used dopants such as group V elements (As and P) and Cu. The carrier densities closely align with experimental results, as indicated in Table 3. It is noteworthy that, based on predicted carrier density, Cu appears to be a more effective dopant. However, experimental data suggests that Cu doping is less effective when compared to group V species. This phenomenon may be attributed to the instability of copper dopants and the presence of compensating defects.[29] Additionally, it is important to acknowledge that the "frozen-in" approximation might not hold in certain cases, particularly when the interactions between different defects are significant, like vacancy-interstitial pair annihilation and exchange reactions.[30] Also, the strong Coulomb interaction between defects with opposite charges can give rise to defect complexes,[2,31] a consideration not addressed within the current framework.

Additionally, we have also identified experimental evidence supporting the efficacy of fluorine doping in CdTe, which aligns with our results and suggests fluorine as a promising dopant, increasing carrier density to around 10^{17} cm^{-3}.[32] Furthermore, as research groups show growing interest in n-type CdTe solar cells,[33] we have also identified potential dopants in Table SIII. Group VII elements are predicted to be effective for rendering CdTe n-type, potentially achieving carrier densities exceeding 10^{15} cm^{-3}. However, the possibility of introducing deep-level traps by these species may impact device performance.

RESULTS AND DISCUSSION

Based on the parity plots of Figure 1 and the columns of Tables 1 and 2 that describe the testing prowess of the models in predicting the two types of formation energies and six types

https://doi.org/10.1021/acs.jpcc.4c01124
J. Phys. Chem. C 2024, 128, 8821−8829

of transition levels, it is observed that the output RMSE performance errors for all types of transition levels are lower than the formation energies. A well-known principle that can explain why predicting transition levels, such as reaction barriers or activation energies, is often easier than predicting formation energies is that they depend on the relative energies of the given states of the local, isolated process within the system rather than considering the total energies of all particles that composed the entire system.

From the formula describing the calculations for the transition levels within the semiconductor (see experimental procedures[12]), such levels can be represented as the energy differences between two charge states. Because these differences depend on the relative positions of the two states in energy space, they can be more accurately predicted than the formation energies, which depend on each particle that makes up the entire structure. Furthermore, it can be shown from the derivation of the transition levels that the chemical potential variable is independent when calculating any defined level compared to the formation energies. This suggests that predicting the formation energies to get the same errors as the transition level requires more features to better incorporate the chemical potential into the network.

Compared to its binary group counterparts, the ANN demonstrates superior prediction accuracy for all output DFT-computed properties in binary group II−VI semiconductors, with smaller errors observed. This enhanced performance can be attributed, first, to the larger portion of II−IV semiconductor data points available for training. However, it is important to consider that the network's underperformance for the IV−IV group can be attributed to the fact that certain crystal compounds within this group may not be stable in their cubic phases (e.g., SiGe). Consequently, this instability leads to unrealistic defect formation energy calculations, expanding the formation energy range and rendering prediction more challenging. As II−VI binary group semiconductors are the most common types of semiconductor system utilized in thin-film solar energy technologies, these low error thresholds suggest that the network is a great tool in predicting the development of new dopants in newer semiconductors for such applications.

The improvements in the performance of ANN and CDGNN models due to the utilization of the SOAP descriptors can be attributed to the fact that such features assist in adding additional information regarding the local defect environment, which describes the region where electrons or holes are exchanged from the atomic impurity to the conduction or valence band, therefore can assist in predicting the electronic properties of defects within the system. Using the Gaussian-smeared atomic density around each atom, the descriptors help catalog the overlaps of electron clouds and the interactive effects of chemical bonding. At the same time, these descriptors are invariant to the permutations of atomic indices, meaning that one can fine-tune the model to help consider such invariance from the material structure without the input of a significant amount of training data to learn such cases, allowing for faster computational times and reducing cost for our neural network framework.

The relatively low error of ANN for II−IV group compound prediction makes quantitative screen and prediction possible. The ML-assisted framework for device design presented in this work is a powerful tool that leverages ML techniques to identify potential dopants to optimize semiconductor device performance. The framework begins with a comprehensive database of defect properties, including defect formation energies and thermodynamic transition levels, generated using DFT and ANN predictions. However, several challenges and limitations must be considered. Notably, while the ML model is capable of predicting a wide range of properties and can optimize device design, it faces challenges when estimating the deep-level capture rate for defects. Accurate predictions for deep-level defects necessitate theoretical calculations and experimental evidence,[42] which the ML model cannot provide at current stage. Additionally, the model's consideration is primarily focused on substitutional defects and interstitials, overlooking other critical point defect types, such as defect complexes,[43,44] which can significantly impact semiconductor devices. To enhance the framework, a more comprehensive defect data set is required. The other limitation is that the data set does not include hexagonal phase semiconductors, such as CdSe and GaN. The framework's performance and applicability are limited to the semiconductor materials present in the available data. Expanding the data set to include hexagonal phase semiconductors could enhance the framework's versatility.

CONCLUSIONS

The ML-assisted device design framework presents a powerful approach to streamline the dopant screening process, resulting in significant time and cost savings. Instead of relying solely on exhaustive experimental trials, the framework utilizes ML to predict defect properties for various dopants across the chemical space. This predictive capability enables researchers to identify promising potential dopants more efficiently, reducing the need for extensive and resource-intensive experimentation.

By employing the comprehensive defect property database generated through DFT, the framework significantly narrows down the pool of potential dopants with desired properties. This targeted selection process minimizes the need to perform numerous experiments, saving both time and labor.

Moreover, the integration of TCAD simulation further refines the search for dopants, focusing on those that have the most favorable characteristics for device optimization. By combining ML predictions with physical models in TCAD simulations, researchers can efficiently gather figures of merit and assess the potential impact of selected dopants on device performance.

The ML-assisted device design framework represents a significant departure from traditional dopant screening methods, providing an innovative, data-driven approach that streamlines the search for superior dopants in semiconductor devices. This transformative framework optimizes resources and expedites the discovery process. Nevertheless, some challenges remain, particularly in accurately predicting deep level capture rates and addressing the absence of defect complexes and hexagonal phase semiconductor data. Furthermore, there is room for improvement in achieving quantitative analysis for III−V and IV−IV types, which might require additional data or more refined descriptors. Future research efforts should be directed toward overcoming these limitations while further advancing the application of this framework in semiconductor device design.

https://doi.org/10.1021/acs.jpcc.4c01124
J. Phys. Chem. C 2024, 128, 8821−8829

■ ASSOCIATED CONTENT

Data Availability Statement

DFT data and neural network algorithms are made available as an open source tool to predict and calculate the neutral formation energies and charge-dependent transition levels of atomic defects and impurities within zinc blende semiconductor systems: https://github.com/dms46/nn_semiconductors/.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcc.4c01124.

Additional simulation details and ML-assisted dopant screening in CdTe (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Xiaofeng Xiang − *Molecular Engineering & Sciences Institute, University of Washington, Seattle, Washington 98195, United States*; orcid.org/0000-0001-7865-1741; Email: xiaofx2@uw.edu

Scott Dunham − *Department of Electrical and Computer Engineering, University of Washington, Seattle, Washington 98195, United States*; Email: dunham@uw.edu

Author

Dylan Soh − *Department of Physics, University of Washington, Seattle, Washington 98195, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcc.4c01124

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Solar Energy Technology Office Award Number DE-EE0008556 and by the National Science Foundation (NSF) MRSEC DMR-1719797 and NSF MRSEC DMR-2308979. We would like to acknowledge the utilization of the Hyak supercomputer system at the University of Washington, supported through CEI, MEM-C, and the Student Technology Fund, which facilitated our research efforts. Furthermore, we extend our gratitude to Professor Arun Mannodi-Kanakkithodi at Purdue University for providing the DFT chemical potential data set and engaging in valuable discussions regarding ML models.

■ REFERENCES

(1) Sommer, D. E.; Dunham, S. T. Atomistic models of Cu diffusion in CuInSe2 under variations in composition. *J. Appl. Phys.* **2018**, *123*, 115116.

(2) Xiang, X.; Sommer, D. E.; Gehrke, A.; Dunham, S. T. Coupled process and device modeling of Cu (In, Ga) Se 2 solar cells. *2021 IEEE 48th Photovoltaic Specialists Conference (PVSC)*, 2021; pp 1707−1711.

(3) Lany, S.; Zunger, A. Accurate prediction of defect properties in density functional supercell calculations. *Modell. Simul. Mater. Sci. Eng.* **2009**, *17*, 084002.

(4) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.

(5) Park, C. W.; Wolverton, C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Phys. Rev. Mater.* **2020**, *4*, 063801.

(6) Choudhary, K.; DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Comput. Mater.* **2021**, *7*, 185.

(7) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **2019**, *31*, 3564−3572.

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

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

(10) 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. *Semicond. Sci. Technol.* **2016**, *31*, 083002.

(11) Mosquera-Lois, I.; Kavanagh, S. R.; Walsh, A.; Scanlon, D. O. Identifying the ground state structures of point defects in solids. *npj Comput. Mater.* **2023**, *9*, 25.

(12) Mannodi-Kanakkithodi, A.; Xiang, X.; Jacoby, L.; Biegaj, R.; Dunham, S. T.; Gamelin, D. R.; Chan, M. K. Y. Universal machine learning framework for defect predictions in zinc blende semiconductors. *Patterns* **2022**, *3*, 100450.

(13) Freysoldt, C.; Grabowski, B.; Hickel, T.; Neugebauer, J.; Kresse, G.; Janotti, A.; Van de Walle, C. G. First-principles calculations for point defects in solids. *Rev. Mod. Phys.* **2014**, *86*, 253−305.

(14) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.

(15) Synopsys Inc.. *Sentaurus Device User Guide Version K-2023.12*; Synopsys, Inc.: Mountain View, CA, 2023.

(16) Himanen, L.; Jäger, M. O.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **2020**, *247*, 106949.

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

(18) Mutter, D.; Dunham, S. T. Calculation of defect concentrations and phase stability in Cu_2ZnSnS_4 and $Cu_2ZnSnSe_4$ from stoichiometry. *IEEE J. Photovolt* **2015**, *5*, 1188−1196.

(19) Britt, J.; Ferekides, C. Thin-film CdS/CdTe solar cell with 15.8% efficiency. *Appl. Phys. Lett.* **1993**, *62*, 2851−2852.

(20) Reich, C. Investigations to Improve CdTe-Based Solar Cell Open Circuit Voltage and Efficiency Using a Passivation and Selectivity Theoretical Framework, Ph.D. Thesis, Colorado State University, 2022.

(21) 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* **2013**, *535*, 322−325.

(22) Yang, J.-H.; Park, J.-S.; Kang, J.; Metzger, W.; Barnes, T.; Wei, S.-H. Tuning the Fermi level beyond the equilibrium doping limit through quenching: The case of CdTe. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *90*, 245202.

(23) Kavanagh, S. R.; Walsh, A.; Scanlon, D. O. Rapid recombination by cadmium vacancies in CdTe. *ACS Energy Lett.* **2021**, *6*, 1392−1398.

(24) Yang, J.-H.; Shi, L.; Wang, L.-W.; Wei, S.-H. Non-radiative carrier recombination enhanced by two-level process: a first-principles study. *Sci. Rep.* **2016**, *6*, 21712.

(25) Krasikov, D. N.; Scherbinin, A. V.; Knizhnik, A. A.; Vasiliev, A. N.; Potapkin, B. V.; Sommerer, T. J. Theoretical analysis of non-radiative multiphonon recombination activity of intrinsic defects in CdTe. *J. Appl. Phys.* **2016**, *119*, 085706.

(26) Xiang, X.; Tong, Y.; Gehrke, A.; Dunham, S. Point defects in CdTe and CdTeSe alloy: a first principles investigation with DFT+U. *arXiv* **2024**, arXiv:2404.07796v2.
(27) Alkauskas, A.; Yan, Q.; Van de Walle, C. G. First-principles theory of nonradiative carrier capture via multiphonon emission. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *90*, 075202.
(28) Turiansky, M. E.; Alkauskas, A.; Engel, M.; Kresse, G.; Wickramaratne, D.; Shen, J.-X.; Dreyer, C. E.; Van de Walle, C. G. Nonrad: Computing nonradiative capture coefficients from first principles. *Comput. Phys. Commun.* **2021**, *267*, 108056.
(29) Perrenoud, J.; Kranz, L.; Gretener, C.; Pianezzi, F.; Nishiwaki, S.; Buecheler, S.; Tiwari, A. N. A comprehensive picture of Cu doping in CdTe solar cells. *J. Appl. Phys.* **2013**, *114*, 174505.
(30) Fahey, P. M.; Griffin, P.; Plummer, J. Point defects and dopant diffusion in silicon. *Rev. Mod. Phys.* **1989**, *61*, 289−384.
(31) Xiang, X.; Sommer, D. E.; Gehrke, A.; Dunham, S. T. Coupled Process/Device Modeling and Point Defect Engineering of Cu-(In,Ga)Se2 Solar Cells. *IEEE J. Photovolt* **2024**, *14*, 422−432.
(32) Ojo, A.; Dharmadasa, I. The effect of fluorine doping on the characteristic behaviour of CdTe. *J. Electron. Mater.* **2016**, *45*, 5728−5738.
(33) Palekis, V.; Wang, W.; Elahi, S. T.; Zahangir Alom, M.; Ferekides, C. Thin Film Solar Cells with n-type CdTe Absorber and p-type ZnTe Window Layers. *2021 IEEE 48th Photovoltaic Specialists Conference (PVSC)*, 2021; pp 1293−1297.
(34) McCandless, B.; Metzger, W. K.; Buchanan, W.; Sriramagiri, G.; Thompson, C.; Duenow, J.; Albin, D.; Jensen, S. A.; Moseley, J.; Al-Jassim, M. Enhanced p-type doping in polycrystalline CdTe films: deposition and activation. *IEEE J. Photovolt* **2019**, *9*, 912−917.
(35) Zhao, Y.; Boccard, M.; Liu, S.; Becker, J.; Zhao, X.-H.; Campbell, C. M.; Suarez, E.; Lassise, M. B.; Holman, Z.; Zhang, Y.-H. Monocrystalline CdTe solar cells with open-circuit voltage over 1 V and efficiency of 17%. *Nat. Energy* **2016**, *1*, 16067.
(36) Oehling, S.; Lugauer, H.; Schmitt, M.; Heinke, H.; Zehnder, U.; Waag, A.; Becker, C.; Landwehr, G. p-type doping of CdTe with a nitrogen plasma source. *J. Appl. Phys.* **1996**, *79*, 2343−2346.
(37) Nagaoka, A.; Nishioka, K.; Yoshino, K.; Katsube, R.; Nose, Y.; Masuda, T.; Scarpulla, M. A. Comparison of Sb, As, and P doping in Cd-rich CdTe single crystals: Doping properties, persistent photoconductivity, and long-term stability. *Appl. Phys. Lett.* **2020**, *116*, 132102.
(38) Nagaoka, A.; Kuciauskas, D.; Scarpulla, M. A. Doping properties of cadmium-rich arsenic-doped CdTe single crystals: Evidence of metastable AX behavior. *Appl. Phys. Lett.* **2017**, *111*, 232103.
(39) Ablekim, T.; Swain, S. K.; Yin, W.-J.; Zaunbrecher, K.; Burst, J.; Barnes, T. M.; Kuciauskas, D.; Wei, S.-H.; Lynn, K. G. Self-compensation in arsenic doping of CdTe. *Sci. Rep.* **2017**, *7*, 4563.
(40) Khan, I. S.; Evani, V. K.; Palekis, V.; Ferekides, C. Effect of stoichiometry on the lifetime and doping concentration of polycrystalline CdTe. *IEEE J. Photovolt* **2017**, *7*, 1450−1455.
(41) Li, D.-B.; Bista, S. S.; Song, Z.; Awni, R. A.; Subedi, K. K.; Shrestha, N.; Pradhan, P.; Chen, L.; Bastola, E.; Grice, C. R.; et al. Maximize CdTe solar cell performance through copper activation engineering. *Nano Energy* **2020**, *73*, 104835.
(42) Zhao, J. H.; Schlesinger, T.; Milnes, A. Determination of carrier capture cross sections of traps by deep level transient spectroscopy of semiconductors. *J. Appl. Phys.* **1987**, *62*, 2865−2870.
(43) Stokes, A.; Al-Jassim, M.; Diercks, D. R.; Egaas, B.; Gorman, B. 3-D point defect density distributions in thin film Cu (In, Ga) Se2 measured by atom probe tomography. *Acta Mater.* **2016**, *102*, 32−37.
(44) Xiang, X.; Gehrke, A.; Dunham, S. Understanding the Dopability of As in Selenium-Alloyed Cadmium Telluride Solar Cells. *2023 IEEE 50th Photovoltaic Specialists Conference (PVSC)*, 2023; pp 1−3.

https://doi.org/10.1021/acs.jpcc.4c01124
J. Phys. Chem. C 2024, 128, 8821−8829