Universal Machine Learning Framework for Defect Predictions in Zinc Blende Semiconductors

Arun Mannodi-Kanakkithodi (✉ amannodi@purdue.edu)

Purdue University

Xiaofeng Xiang

Molecular Engineering & Sciences Institute, University of Washington <https://orcid.org/0000-0001-7865-1741>

Laura Jacoby

Department of Chemistry, University of Washington

Robert Biegaj

Materials Science & Engineering, University of Washington

Scott Dunham

Department of Electrical and Computer Engineering, University of Washington

Daniel Gamelin

Department of Chemistry, University of Washington

Maria Chan

Center for Nanoscale Materials, Argonne National Laboratory <https://orcid.org/0000-0003-0922-1363>

Article

Keywords: density functional theory (DFT), semiconductors, condensed matter physics

Posted Date: July 26th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-723035/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Patterns on February 1st, 2022. See the published version at <https://doi.org/10.1016/j.patter.2022.100450>.

Universal Machine Learning Framework for Defect Predictions in Zinc Blende Semiconductors

Arun Mannodi-Kanakkithodi*,^{1,2,a)} Xiaofeng Xiang*,³ Laura Jacoby*,⁴ Robert Biegaj,⁵ Scott T. Dunham,⁶ Daniel R. Gamelin,⁴ and Maria K.Y. Chan^{1,b)}

¹⁾*Center for Nanoscale Materials, Argonne National Laboratory, Argonne, Illinois 60439, USA*

²⁾*School of Materials Engineering, Purdue University, West Lafayette, Indiana 47907, USA*

³⁾*Molecular Engineering & Sciences Institute, University of Washington, Seattle, Washington 98195, USA*

⁴⁾*Department of Chemistry, University of Washington, Seattle, Washington 98195, USA*

⁵⁾*Materials Science & Engineering, University of Washington, Seattle, Washington 98195, USA*

⁶⁾*Department of Electrical and Computer Engineering, University of Washington, Seattle, Washington 98195, USA*

ABSTRACT

Point defects or impurities are either naturally present in semiconductors or may be intentionally introduced to tune their electronic and optical properties. The nature of impurity energy levels can strongly influence the performance of a semiconductor in applications ranging from solar cells to photodiodes to infrared sensors to qubits for quantum computing. In this work, we develop a framework powered by machine learning (ML) and high-throughput density functional theory (DFT) computations for the prediction and screening of functional impurities in group IV, III-V, and II-VI zinc blende semiconductors. Elements spanning the length and breadth of the periodic table are considered as impurity atoms at the cation, anion, or interstitial sites in supercells of 34 candidate semiconductors, leading to a chemical space of $\sim 12,000$ points, 10% of which are used to generate a DFT dataset of charge dependent defect formation energies. Descriptors based on tabulated elemental properties, defect coordination environment, and relevant semiconductor properties are used to train ML regression models for the DFT computed properties, resulting in statistical predictions of the neutral state formation energies and charge transition levels of all possible impurities in the given set of compounds. Kernel ridge regression, Gaussian process regression, and neural networks, with appropriate feature selection and hyperparameter optimization, are seen to yield similar predictive performances and meaningful uncertainty estimates. We apply the ML framework to screen all impurities with lower formation energy than dominant

native defects in all group IV, III-V, and II-VI zinc blende semiconductors. An online tool resulting from this work for predicting and visualizing defect properties in semiconductors is made available on github.

INTRODUCTION

Compositional manipulation of semiconductors is one of the primary methods utilized to obtain optimal properties^{1–6}. Apart from alloying, the primary means for compositional control of semiconductor properties is the introduction of dopants or impurities, i.e. guest atoms at a cation, anion, or interstitial sites. Such impurities, even in a very dilute concentration, can potentially cause major changes in the electronic structure and physical properties of the material^{7–10}. The complete understanding of a semiconductor’s optoelectronic behavior requires estimating the formation energies of point defects, whether accidental or intentionally introduced^{6,11,12}.

While $\sim 90\%$ of solar cells still rely on crystalline Si as the absorber, related group IV semiconductors such as SiC, II-VI semiconductors such as CdTe, III-V semiconductors such as GaAs, and various derivative compounds are all viable as photovoltaic (PV) materials and are currently in use in single terminal as well as tandem solar cells^{2,13–17}. Many of these compounds have also been used in transistors, photodiodes, lasers, and qubits or quantum sensors. The chemical space of binary group IV, III-V, and II-VI semiconductors contains compounds that exist in the cubic zinc blende (ZB) or wurtzite crystal structures and show systematic trends in lattice constants, electronic band gaps, optical absorption coefficients, and defect properties¹⁸. Alloying in these spaces has frequently been used for tuning properties and performance, with some prominent examples including the use of CdSeTe in solar cells^{19,20} and AlGaAs in LEDs^{21,22}.

^{a)}Electronic mail: amannodi@purdue.edu

^{b)}Electronic mail: mchan@anl.gov

FIG. 1: The semiconductor-impurity chemical space in terms of the cation (A) and anion (B) choices for group IV, II-VI, and III-V compounds, types of defect sites, and impurity atoms selected from across the periodic table.

Although the structure and optoelectronic properties of binary, ternary and even quaternary compounds in the group IV, III-V, and II-VI semiconductor space have been widely studied both computationally and experimentally^{2,3,11,18}, a comprehensive understanding of the formation likelihood and electronic levels of point defects and impurities is missing. A look at functional atomic defects in semiconductors reveals that the energy levels created inside the band gap can (a) reduce PV efficiency via nonradiative recombination of charge carriers, (b) enable sub-gap absorption or emission if the levels are partially filled or if they have low photoionization energies, and (c) enable quantum computing, quantum sensing, and quantum communication via their nuclear or electronic spins. A universal prediction framework for impurity behavior in known and novel semiconductor spaces is thus paramount¹¹. Given such a framework for group IV, III-V, and II-VI semiconductors, it would be possible to perform high-throughput screening of impurity atoms from across the periodic table in terms of their energetics relative to dominant native defects (such as vacancies and self-interstitials), nature of equilibrium conductivity, and location of energy levels with respect to band edges.

For years, defect levels and their donor or acceptor type nature have been experimentally measured using methods such as deep level transient spectroscopy (DLTS) and cathodoluminescence (CL)^{23,24}, but such studies have been limited by difficulties in sample preparation and assigning measured levels to specific vacancies, interstitials, substitutions, or complex defects. Computationally, first principles density functional theory (DFT) has been widely used to predict the formation

energies of point defects as a function of the net charge in the system, the chemical potential conditions, and the Fermi level as it goes from the valence band maximum (VBM) to the conduction band minimum (CBM)^{6,10–12,25}. When an appropriate level of theory is applied, DFT computed defect charge transition levels have been seen to match well with measured levels and have helped identify specific charge transitions of specific defects. DFT can reliably predict defect and impurity behavior in a variety of semiconductors, but limitations arise from the computational expense of using large supercells and performing charged calculations¹¹, making it difficult to extend calculations to explore new systems broadly.

Predictive machine learning (ML) models, trained from existing or freshly generated data, act as surrogates for DFT calculations by providing statistical estimates of the desired properties^{11,26–28}. The burgeoning field of materials informatics has led to many successes, with some of the most notable contributions resulting from the combination of first principle computations and machine learning. ML applied on DFT data has seen the development of predictive and design tools^{29–31}, the discovery of novel materials for batteries, capacitors, solar cells, and thermoelectrics^{32–36}, and the efficient exploration of extremely large chemical spaces^{37,38}. Recent work from our group involved performing high-throughput DFT computations to study the formation energies and charge transition levels of impurities in halide perovskites⁶ and Cd-chalcogenides¹¹, following which ML models were trained for the prediction and screening of impurity atoms that can shift the equilibrium Fermi level as determined by dominant native defects. An extension of these studies in terms of semi-

conductor and impurity chemical spaces as well as ML techniques can pave the path towards a universal framework for impurity prediction and design.

In this work, we consider atomic impurities from across the periodic table, in a chemical space of binary group IV, III-V, and II-VI semiconductors in the ZB structure, and use the DFT+ML methodology to predict their complete charge, chemical potential, and Fermi level dependent formation energies. This is a direct extension of our work on Cd-chalcogenides¹¹, which forms a subset of the computational data presented here. We perform high-throughput DFT computations on impurity atoms simulated at the cation, anion, and different interstitial sites in several selected compounds in the group IV (e.g. Si, SiC, GeC, etc.), III-V (e.g. BSb, GaAs, InP, etc.), and II-VI (e.g. ZnSe, CdS, etc.) chemical space, and use descriptors encoding information about the semiconductor, the impurity atom, and the defect site coordination environment as input to train ML models that predict the neutral state formation energy and 6 types of charge transition levels for any possible impurity. We used sure independence screening and sparsifying operator (SISSO) for feature selection and K-nearest neighbors (KNN) approach for outlier detection, followed by regression techniques such as random forest, Gaussian process, and neural network to yield the predictive models.

In the following sections, we discuss the exact composition of the chemical space and visualize the DFT computed data, also plotting the Fermi level dependent formation energies of native defects and impurities for selected compounds. We then delve deep into the development of the machine learning framework, explaining the descriptor choices, methods of feature selection, outlier exclusion, and the various regression techniques used. We compare the performances of different models using root means square errors (RMSE) and estimate the uncertainties in prediction for each technique. The best models thus obtained are used to make predictions for the entire chemical space, only $\sim 10\%$ of which was used to generate the DFT data, and make a list of "dominating" impurities for each compound. We finish with a perspective on what can be accomplished using this design framework, the limitations of this work, and potential next steps. The DFT data and ML codes generated through this work are made available on github.

RESULTS AND DISCUSSION

Semiconductor and Impurity Chemical Space

The chemical space considered in this work has been pictured in Fig. 1 in terms of the semiconductor compounds, possible defect sites, and impurity atoms. We include AB semiconductors (with A broadly defined as the cation and B the anion) belonging to groups II-VI, III-V and IV-IV, leading to 8 II-VI compounds (CdO, CdS, CdSe, CdTe, ZnO, ZnS, ZnSe, and ZnTe), 16 III-V compounds (BN, BP, BAs, BSb, AlN,

AlP, AlAs, AlSb, GaN, GaP, GaAs, GaSb, InN, InP, InAs, and InSb), and 10 group IV compounds (C, Si, Ge, Sn, SiC, GeC, SnC, SiGe, SiSn, and GeSn). The resulting 34 compounds are modeled in the cubic ZB structure, with A atoms occupying an FCC lattice and B atoms occupying the tetrahedral sites. The DFT computed lattice constants and band gaps (using different levels of theory) of all the compounds are listed in Table ??, along with corresponding experimental measurements collected from the literature. It can be seen that while the cubic lattice constants are reasonably accurate, the standard GGA-PBE functional underestimates the band gap, as has been well demonstrated in the past³⁹⁻⁴¹. Band gaps computed for some compounds using the hybrid HSE06 functional, with and without spin-orbit coupling (SOC), compare better with experiments.

In any AB compound in the ZB structure, defects or impurities could be found at the A-site, B-site, or several possible symmetrically-inequivalent interstitial sites. Fig. 1 also shows the defect sites considered in this work, namely the A and B sites and 3 types of interstitial sites: the A-site interstitial (with 4 neighboring A atoms), the B-site interstitial (with 4 neighboring B atoms), and neutral site interstitial (with 3 neighboring A and B atoms each). The 5 defect sites are considered in the 30 binary compounds while in the remaining 4 elemental systems (C, Si, Ge, Sn), 3 defect sites are considered (A-site, A-site interstitial and neutral site interstitial). In terms of impurity atoms, we consider nearly all elements from periods II to VI as well as all lanthanides, leading to a total of 77 species, as pictured in Fig. 1. The total number of possible impurities in this chemical space can thus be estimated as: $77 \times 5 \times 30 + 77 \times 3 \times 4 = 12474$. Out of these 12474 data points, about 10% are considered for DFT computations to determine their neutral state formation energies and charge transition levels; ML models trained on these data based on the properties of the semiconductor compound, defect site coordination, and impurity atom lead to generalized predictions applicable to all the data points.

Defect Formation Energy

All native defects and impurities were simulated in 64 atom $2 \times 2 \times 2$ supercells of the parent compound, based on previously optimized 8 atom ZB unit cells. DFT optimization was performed in the neutral and charged states ($q = -3, -2, -1, 0, +1, +2, +3$) while keeping the supercell shape and size fixed. All computations were performed using the Vienna ab-initio Simulation Package (VASP)^{58,59} employing the Perdew-Burke-Ernzerhof (PBE)⁶⁰ exchange-correlation functional and projector-augmented wave (PAW) atom potentials⁶¹. The kinetic energy cut-off for the planewave basis set was 500 eV, and all atoms were relaxed until forces on each were less than 0.05 eV/Å. Brillouin zone integration was performed using a $3 \times 3 \times 3$ Monkhorst-Pack mesh. For any defect or impurity atom M in a compound AB, the following equations yield the formation energy E^f as a function of the chemical potential μ ,

FIG. 2: Comparison of DFT computed defect levels with experimentally measured levels (obtained from publications^{42–57}). Measured vs DFT RMSE values are also shown for different semiconductor types and for the combined set of points. A few defect levels have been labeled.

charge q , and Fermi level E_F , and any impurity charge transition level, $\varepsilon(q_1/q_2)$:

$$E^f(q, E_F) = E(M^q) - E(AB) + \Delta\mu + q(E_F + E_{vbm}) + E_{corr} \quad (1)$$

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

Here, $E(AB)$ is the DFT energy of an AB supercell without defects, $E(M^q)$ is the DFT energy of the AB supercell containing a defect M in a charge state q , E_{vbm} is the valence band maximum as computed from an electronic structure calculation on AB, and E_{corr} is the charge correction energy using the scheme developed by Freysoldt et al.^{62,63} to account for periodic interaction between image charges. E^f depends on the chemical potential change $\Delta\mu$ involved in creating the defect, and for a given $\Delta\mu$, it is a function of E_F and q such that the slope of the E^f vs E_F plot is equal to q . For any defect or impurity M in compound AB, the chemical potentials of all species are defined with reference to the elemental standard states of M, A, and B, as well as their lowest formation energy binary or ternary compounds. For an impurity M_A (M occupying an A site), in Eqn. 1, $\Delta\mu = \mu_A - \mu_M$; for an impurity M_i (M occupying an interstitial site), $\Delta\mu = -\mu_M$; for a vacancy at the B site V_B , $\Delta\mu = \mu_B$. We calculate formation energies at two extreme chemical potential conditions, namely A-rich (where

μ_A = energy of elemental standard state of A) and B-rich (where μ_B = energy of elemental standard state of B), and note that by tuning the μ conditions, defects can be made more or less stable, and the equilibrium conductivity—determined by defect charge neutrality conditions—can be made more p-type or n-type. Eqn. 2 defines a charge transition level $\varepsilon(q_1/q_2)$, i.e. the E_F value where the defect transitions from a charge state q_1 to q_2 , which is independent of the μ conditions; in this work, for every defect or impurity, we calculate six possible transition levels, namely $+3/+2$, $+2/+1$, $+1/0$, $0/-1$, $-1/-2$, and $-2/-3$.

An experimental comparison of defect properties computed at the chosen level of theory is worthwhile prior to launching a computational data-driven discovery exercise. As noted earlier, the standard GGA-PBE functional used in this work is known to underestimate band gaps, but it has been reported that defect charge transition levels from PBE can compare well with experiments for various semiconductor classes such as halide perovskites^{6,9} and group IV, III-V, and II-VI semiconductors^{12,25,64}. This contrast can be attributed to total energy differences in DFT being more accurate than using Kohn-Sham energy levels to estimate band edges^{12,25,64} or band gaps³⁹, or in the case of halide perovskites, a fortuitous cancellation of errors that leads to PBE being as accurate as HSE06+SOC^{6,9}. To ascertain the accuracy of PBE defect and impurity levels in all II-VI, III-V, and IV-IV compounds, we scoured the published literature^{42–57} (in a brute-force manner, something we hope to replace with more efficient and comprehensive natural language processing-based searches in the future^{65,66}) and collected measured energy levels for 84 defects across the 34 compounds, adding to the set of 15 points collected for CdTe in reference¹². As presented in Fig. 2 and Table ??, the PBE predicted defect levels are highly correlated with experimental measurements, with a correlation coefficient (R^2) of 0.85. We find that the root mean square error (RMSE) of PBE predictions compared to experiments is below 0.2 eV for III-V and IV-IV compounds, and 0.24 eV for II-VI compounds, resulting in a total RMSE of around 0.21 eV. This is an acceptable level of accuracy that is similar to what we found in reference¹², and is within the recognized accuracy limit of DFT electronic levels; a similar ML vs DFT accuracy would be desired for eventual prediction and screening to be performed with some degree of experimental precision. To our knowledge, this is the largest comparison performed to date between DFT computed defect levels and experimental measurements.

Before discussing the computational dataset of impurity formation energies and charge transition levels, we take a look at the complete picture of native point defect formation energies as a function of q , μ , and E_F . In each of the 34 semiconductors, we performed neutral and charged DFT calculations for all possible vacancy, interstitial, and anti-site substitutional defects. It should be noted that all interstitial and substitutional native defects are made a part of the impurity DFT data for machine learning but vacancies are not, as in the current ML framework, many descriptor dimensions are made up of properties of the atom occupying the defect site. In Fig. 3,

FIG. 3: Computed formation energies of native defects (solid lines) and selected impurities (dashed lines) in (a) ZnSe under Se-rich conditions, (b) AlAs under As-rich conditions, and (c) SiC under Si-rich conditions, as a function of the Fermi level as it goes from the VBM ($E_F = 0$ eV) to the CBM ($E_F = \text{experimental band gap}$). The intersection point of the dominant donor and acceptor type native defects (shown using extended dotted colored lines) approximately gives the equilibrium defect formation energy, and the vertical dotted lines show the equilibrium Fermi level. Some charge transition levels and neutral state formation energies have been labeled.

we plotted the computed formation energies as a function of Fermi level (as it goes from the VBM to the CBM) for all native point defects and selected impurities in (a) ZnSe at Se-rich conditions, (b) AlAs at As-rich conditions, and (c) SiC at Si-rich conditions.

From Fig. 3, we can deduce the lowest energy donor and acceptor type native defects, their preferred charged states inside the band gap, the p-type or n-type nature of equilibrium conductivity, and the energetics of impurities relative to dominant native defects. For instance, in ZnSe at Se-rich conditions, V_{Zn} and Zn_i are the dominant acceptor and donor type defects respectively, and pin the equilibrium E_F (determined by applying charge neutrality conditions⁶⁷) closer to the valence band edge, indicating a p-type conductivity. It can be seen that impurities Pt_{Zn} and Cl_i create higher energy negatively charged defects in the band gap than V_{Zn} , meaning they cannot compensate for native defects. Similarly, V_{Al} and As_{Al} form the lowest energy acceptor and donor type defects in AlAs at As-rich conditions and pin the equilibrium E_F around the middle of the band gap, resulting in an intrinsic type conductivity, while impurities Ti_{As} and C_i are both higher energy defects. In SiC at Si-rich conditions, V_{C} and C_{Si} are the lowest energy donor and acceptor type defects and lead to intrinsic conductivity. Also marked in Fig. 3 (a) to (c) are some neutral state formation energies, ΔH , and some charge transition levels, $\epsilon(q_1/q_2)$. In Table ?? and ??, we list the dominating acceptor and donor type defects, and the equilibrium E_F , E^f , and type of conductivity in every compound at A-rich and B-rich chemical potential conditions; all impurity formation energies will be compared against these defects to determine whether they are dominating or not, and how they might change the equilibrium conductivity.

Computational Dataset

Building upon the dataset of impurity properties in Cd-chalcogenides from our previous work¹² and the native defects presented in the previous section, we performed additional impurity calculations for randomly selected impurity atoms across the space of 34 group IV, III-V, and II-VI semiconductors. For any defect or impurity, what is ultimately desired is a complete picture of formation energy as a function of charge, Fermi level, and chemical potential, as shown in Fig. 3. We explicitly considered two types of predictable properties which can yield the entire formation energy picture, namely the neutral state formation energy (ΔH (A-rich) or ΔH (B-rich)) and possible charge transition levels ($\epsilon(+3/+2)$, $\epsilon(+2/+1)$, $\epsilon(+1/0)$, $\epsilon(0/-1)$, $\epsilon(-1/-2)$ and $\epsilon(-2/-3)$). Out of a total chemical space of 12,474 impurity types, we computed using DFT (a) 1568 ΔH values at either chemical potential condition, and (b) 1004 $\epsilon(q_1/q_2)$ values for all 6 charge transition types. We then train eight separate ML models for ΔH (A-rich), ΔH (B-rich), and transition levels from $\epsilon(+3/+2)$ to $\epsilon(-2/-3)$.

In Fig. ??, the distributions of semiconductor types (II-VI, III-V, or IV-IV) and impurity types (A-site, B-site, or interstitial site) are pictured for the entire chemical space and for the two DFT datasets. It can be seen that based on the chemical space we selected, almost half the data points belong to III-V semiconductors and a quarter of the points each belong to II-VI and IV-IV; this is, however, not reflected in the DFT datasets, owing to a predominance of data available on II-VI semiconductors from past and present work. Nevertheless, it is expected that the III-V and IV-IV semiconductors are adequately represented because while choosing data points for DFT calculations, it was ensured that at least 10 impurities in each compound are selected, and every defect site is considered. Further, the entire chemical space contains about 40% substitutional impurities and 60% interstitial impurities, with the

FIG. 4: Visualization of DFT data: (a) to (c) transition levels (+3/+2) to (-2/-3), and (d) neutral state formation energies at A-rich and B-rich chemical potential conditions.

former being equally divided between A-site and B-site substitutions, and the latter divided equally between three types of interstitial sites; all the defect sites are pictured in an example ZB supercell in Fig. 1. The defect site distribution is similar for each defect property (ΔH or ε), ensuring adequate representation.

To visualize the DFT data, we plotted all the computed charge transition levels and formation energies in Fig. 4. The transition levels are plotted two at a time against each other in Fig. 4 (a), (b), and (c), as a fraction of the experimental band gap of the compound. Many transition levels are seen to lie deep inside (> 0.2 eV from the band edges) the shaded region that represents the band gap, which indicates the tendency of certain impurities to create deep energy levels. It can also be seen that most of the mid-gap impurity levels belong to +1/0 and 0/-1 transitions, and to a lesser extent to +2/+1 and -1/-2, but almost not at all to the higher charge transitions like +3/+2 and -2/-3. The ranges of values of the transition levels are fairly wide, from deep inside the VB to the band gap to deep inside the CB, which reveals a great variety in the type of impurities based on their preferred oxidation states and the sites they occupy. In Fig. 4 (d), we plotted ΔH (A-rich) vs ΔH (B-rich), which shows values that range from ~ -5 eV to ~ 20 eV. For group IV compounds C, Si, Ge and Sn, A-rich and B-rich conditions are the same, leading to many of the red points lying along the diagonal. In general, ΔH (A-rich) and

ΔH (B-rich) pin two extremes of the impurity formation energy values, and medium chemical potential conditions would lead to intermediate formation energies.

Machine Learning Framework

Descriptors

Aside from generating the computational data, the need for *domain expertise* is most evident in creating appropriate descriptors or sets of features that can uniquely represent every point in the dataset. In the semiconductor+impurity chemical space used in this work, we can uniquely identify every data point using the identity of the semiconductor, the identity of the impurity atom, and the defect site it occupies. Thus, we define descriptors for any impurity M at any site S in any compound AB by combining the following three levels of information:

1. AB_{prop} : Available computed or experimental properties of the semiconductor AB, such as the formation energy, lattice constant, band gap and dielectric constant; this leads to 5 dimensions.
2. $Elem_{prop}$: Tabulated elemental properties of the impu-

rity M as well as species A and B , such as ionic radius, ionization energy, electronegativity, etc.; this leads to 81 dimensions.

3. CM_{prop} : Quantifying the chemical coordination environment around the defect site S in terms of A and B neighbors, using the Coulomb Matrix definition⁶⁸; this leads to 8 dimensions.

The complete list of descriptors can be found on the x-axis of Fig. ?? as well as Table ??, which show the Pearson coefficient of linear correlation ($|r|$) between the properties of interest, ΔH and $\epsilon(q_1/q_2)$, and each of the descriptors.

Feature Selection

The primary feature set of 94 dimensions are all assumed to be relevant to describe the targeted predictors, that is, the impurity transition levels and formation energies. To better explore the nonlinear relationships that may exist between these descriptors dimensions and the properties, we used the sure independence screening and sparsifying operator (SISSO)⁶⁹ method to perform feature engineering. First, a set of operators, namely '+, -, *, /, exp, log, ^(-1), ^2, ^3, sqrt, cbrt, |·|', are implemented recursively for feature space expansion. The total feature size goes from $\sim 10^2$ to $\sim 10^5$ after two iterations. Next, sure independence screening⁷⁰ is used to screen all features from 0D feature space (no iteration) to 1D feature space (one iteration) to 2D feature space (two iterations) using a linear correlation metric, leaving behind only highly correlated features. Finally, a sparsifying operation⁷¹ is applied to filter down the feature space to 80–150 for each output.

Outlier Identification

The detection of outliers in a dataset helps identify candidates with unusual properties which may either be erroneous or lead to lower accuracy when used to train ML models. K-nearest neighbors (KNN) is a method commonly applied to identify the outliers in a dataset based on a feature space⁷². KNN assigns classes to data points based on the most common assignment of its k nearest neighbors; any point that is surrounded by points belonging to a different class is denoted an outlier. Here, the DFT computed transition levels and formation energy values were used as a combined input to a KNN framework. Another method we considered to detect outliers was Principal Component Analysis (PCA)⁷³, which reduces the number of variables used to describe an output while still maintaining most of the descriptive information. A covariance matrix of the data is decomposed to orthogonal eigenvectors, associated with eigenvalues which signify how much of the variance in the data that eigenvector captures. Each data point is labeled with an outlier score as determined by the sum of the weighted distances to all the eigenvectors, with smaller eigenvalues having higher influence (as this is where

outliers are more likely to exist)⁷⁴. Using both KNN and PCA, we selected 10% of the data with the highest outlier scores to be removed. We found that PCA disproportionately removed data points belonging to IV-IV semiconductors. Ultimately, the KNN filtered inlier points proved more effective for all models, and were employed as the standard training sets.

Training Regression Models

In the following sections, we will discuss the optimization and performance of various regression models trained on the computational data, starting with linear regression and moving on to different nonlinear regression techniques, namely random forests, Gaussian processes, kernel ridge, and neural network regression. Common to every method used in this work is the way training-test split, cross-validation, hyperparameter optimization, and error evaluation was performed. A separate model was trained for each of the 8 outputs, namely the 6 types of transition levels and 2 formation energies. Five-fold cross validation was implemented for each model because of a strong dependence of the prediction ability on the exact points chosen for training. Cross-validation helps reduce the reported bias and variance, and is important for avoiding *overfitting*. Various important hyperparameters were optimized for each regression technique; for instance, for neural networks, they include the number of hidden layers, the numbers of nodes in each layer, the dropout rate, etc. All regression models were trained using functions in the Python machine learning library scikit-learn⁷⁵.

The metric for evaluating model performance was chosen to be the prediction RMSE. Each of the 5 folds was treated as a validation set over multiple training cycles, and the prediction RMSE for each fold was averaged over the number of folds. This leads to an effective 80-20 training-test split in the dataset, and an effective "test prediction error" is obtained for every data point, providing an unbiased prediction that reveals the true predictive power of the trained model. The optimal set of hyperparameters is chosen such that the cross-validation error is minimized. Further, the standard deviation in predictions over the multiple training cycles is defined as the uncertainty for each predicted point, providing an error bar that accompanies every prediction. Results are presented as parity plots of ML predictions vs. DFT-calculated properties, with reported RMSE values in eV, and plots between uncertainties and errors in every data point; the predictions are also visualized in terms of semiconductor type. The training prediction RMSE values are listed in Tables ?? and ??, and the test prediction RMSE values listed in Tables I and II, divided in terms of property, ML technique used, and type of data point.

Linear Regression

Fig. ?? shows the Pearson coefficient of linear correlation between various (primary) descriptor dimensions and the prop-

erties of interest. We see that many of the features are 50–70% correlated with the properties, showing a certain degree of linear relationship. To explore this further, we chose multiple linear regression (MLR) as the method to train the first predictive models. Given a vector of properly standardized features $X^T = (X_1, X_2, \dots, X_p)$, and calculated output vector Y , the matrix of coefficients β corresponding to each feature and output is determined by minimizing the least square error $|Y - X^T \beta|^2$. While MLR yields an unbiased predictor, it is prone to overfitting when several features are highly correlated with the output. To address this issue, we use three shrinkage methods, namely LASSO (least absolute shrinkage and selection operator) regression, ridge regression, and elastic net regression⁷⁶, all of which yield a biased predictor but with lower variance, leading to less overfitting compared to standard least square. Ridge regression shrinks the coefficients β by imposing an L_2 penalty whereas LASSO uses an L_1 penalty⁷⁷. Elastic Net is another regularized linear regression technique that combines both L_1 penalty and L_2 penalty. Typically, β shrinkage inside LASSO regression progresses more severely compared to the other two approaches, and some of the coefficient are brought down to 0.

In Figs. ??, ??, ??, and ??, we present parity plots for models trained using MLR, LASSO, ridge regression, and elastic net regression, respectively. We see from the plots and from Tables I and II that there is a marginal improvement in prediction going from MLR to LASSO, ridge, or elastic net regression. The presented results are using the SISSO features as they provide better predictions than using the primary feature set, which is presumably due to the non-linear nature of the input-output relationship. We further find that there is a strong dependence of the prediction error on semiconductor type and impurity site. We observe these effects in non-linear models as well, which will be discussed in subsequent sections. Such prediction error differences can be attributed to the imbalanced distribution of training data; the DFT datasets are biased towards II-VI semiconductors and interstitial site impurities, as seen from Fig. S1. In general, we get better performance on II-VI or interstitial data points since the models we trained work better on majority groups.

Random Forest Regression

The improvement in linear regression model performances upon going from using the primary features to the SISSO-based features shows the importance of interpreting non-linear relationships between the features and properties. However, non-linearity is still limited by the set of operators used in the SISSO method; to further explore this effect, we adopted a popular non-linear regression algorithm known as random forest regression (RFR). RFR is an ensemble measurement method that fits a designated number of classifying decision trees such that each tree is fit on a different randomized subsample of the data set, chosen through bootstrapping. During the construction of any tree, the best split for each node is found based on some number of input features. Averag-

ing over all the trees in the forest can be performed in several ways, and in this work, the model combines the results of the trees by averaging their probabilistic prediction, which improves prediction accuracy and can help control overfitting⁷⁸.

Hyperparameter tuning focuses on the five most important features in the RFR model, namely, the number of trees in the forest, the maximum depth of each tree, the number of features to consider when looking for the best node split, the minimum number of samples required to split an internal node, and the minimum number of samples required to be at a leaf node. For each of the eight outputs, Bayesian optimization was performed⁷⁹ using a function set to minimize both the test RMSE and the difference between the training and test RMSE to balance the bias-variance trade off in the model. Parity plots for the optimized models for all 8 properties are shown in Fig. 5(a). Looking at the error values listed in Tables I and II, there is a general improvement in all the transition level prediction RMSEs from between 0.3 and 0.45 eV for the linear models to between 0.25 to 0.38 eV for RFR, and the formation energy RMSEs drop from 1.2 eV or higher to between 1.07 and 1.16 eV. We further plotted the RFR uncertainties in prediction against the absolute prediction error values in Fig. 5(b). While uncertainty and error do not correlate linearly, such plots reveal the degree of confidence one can have in a given prediction. A large portion of points lie in the low uncertainty, low error region, but a number of points with low uncertainty also show large prediction errors, which highlights the need to exercise caution in trusting the ML predictions regardless of the estimated uncertainty.

As observed in the linear regression results, we found that RFR was able to predict formation energies of II-VI semiconductor impurities more accurately than the III-V or IV-IV, owing to the larger portion of II-VI semiconductor points in the training data set. Interestingly, the transition levels showed much less of a dependence on semiconductor type; the difference could be due to the larger range of values in the formation energy data versus the transitional levels. We found that the points the model predicted most inaccurately for formation energies are relative outliers as predicted by KNN and PCA, and of those points, III-V and IV-IV semiconductor types make up a larger portion than in the data set as a whole. When analyzing the prediction results by site of impurity defect, it was once again seen that interstitials are predicted slightly better than substitutionals, once again owing to the predominance of the former in the dataset.

Kernel Ridge Regression

The improvement in prediction with RFR provided the motivation for alternative nonlinear regression techniques that could lead to further lowering of errors. Kernel ridge regression (KRR) is a similarity-based regression technique which uses the kernel trick to solve a non-linear problem in a linear fashion. The original low-dimensional features are used as input and mapped to a high-dimensional kernel space in which

TABLE I: ML test set prediction RMSE values for transition levels.

Property	ML Method	II-VI Error (eV)	III-V Error (eV)	IV-IV Error (eV)	Total Error (eV)
$\epsilon(+3/+2)$	MLR	0.35	0.37	0.34	0.35
$\epsilon(+3/+2)$	Ridge	0.35	0.35	0.32	0.34
$\epsilon(+3/+2)$	LASSO	0.36	0.36	0.32	0.35
$\epsilon(+3/+2)$	Elastic Net	0.35	0.35	0.32	0.34
$\epsilon(+3/+2)$	RFR	0.36	0.31	0.35	0.34
$\epsilon(+3/+2)$	KRR	0.33	0.37	0.31	0.33
$\epsilon(+3/+2)$	GPR	0.32	0.36	0.32	0.33
$\epsilon(+3/+2)$	NN	0.29	0.36	0.29	0.31
$\epsilon(+2/+1)$	MLR	0.42	0.46	0.46	0.44
$\epsilon(+2/+1)$	Ridge	0.42	0.43	0.45	0.43
$\epsilon(+2/+1)$	LASSO	0.43	0.44	0.45	0.44
$\epsilon(+2/+1)$	Elastic Net	0.42	0.43	0.45	0.43
$\epsilon(+2/+1)$	RFR	0.39	0.36	0.40	0.38
$\epsilon(+2/+1)$	KRR	0.33	0.38	0.40	0.36
$\epsilon(+2/+1)$	GPR	0.32	0.38	0.41	0.36
$\epsilon(+2/+1)$	NN	0.29	0.35	0.38	0.33
$\epsilon(+1/0)$	MLR	0.40	0.39	0.43	0.40
$\epsilon(+1/0)$	Ridge	0.40	0.38	0.42	0.40
$\epsilon(+1/0)$	LASSO	0.41	0.39	0.43	0.41
$\epsilon(+1/0)$	Elastic Net	0.40	0.38	0.42	0.40
$\epsilon(+1/0)$	RFR	0.38	0.36	0.39	0.38
$\epsilon(+1/0)$	KRR	0.31	0.34	0.38	0.33
$\epsilon(+1/0)$	GPR	0.29	0.32	0.38	0.32
$\epsilon(+1/0)$	NN	0.29	0.31	0.37	0.32
$\epsilon(0/-1)$	MLR	0.37	0.42	0.34	0.38
$\epsilon(0/-1)$	Ridge	0.37	0.40	0.34	0.37
$\epsilon(0/-1)$	LASSO	0.37	0.40	0.34	0.37
$\epsilon(0/-1)$	Elastic Net	0.37	0.40	0.34	0.37
$\epsilon(0/-1)$	RFR	0.37	0.33	0.35	0.35
$\epsilon(0/-1)$	KRR	0.32	0.36	0.32	0.33
$\epsilon(0/-1)$	GPR	0.31	0.34	0.32	0.32
$\epsilon(0/-1)$	NN	0.28	0.33	0.31	0.30
$\epsilon(-1/-2)$	MLR	0.33	0.38	0.30	0.33
$\epsilon(-1/-2)$	Ridge	0.32	0.37	0.29	0.32
$\epsilon(-1/-2)$	LASSO	0.32	0.37	0.29	0.33
$\epsilon(-1/-2)$	Elastic Net	0.32	0.37	0.29	0.33
$\epsilon(-1/-2)$	RFR	0.34	0.35	0.27	0.33
$\epsilon(-1/-2)$	KRR	0.29	0.32	0.27	0.29
$\epsilon(-1/-2)$	GPR	0.29	0.31	0.28	0.29
$\epsilon(-1/-2)$	NN	0.26	0.29	0.28	0.27
$\epsilon(-2/-3)$	MLR	0.27	0.26	0.22	0.26
$\epsilon(-2/-3)$	Ridge	0.27	0.26	0.22	0.25
$\epsilon(-2/-3)$	LASSO	0.27	0.26	0.22	0.25
$\epsilon(-2/-3)$	Elastic Net	0.27	0.26	0.22	0.25
$\epsilon(-2/-3)$	RFR	0.24	0.28	0.27	0.25
$\epsilon(-2/-3)$	KRR	0.26	0.24	0.21	0.24
$\epsilon(-2/-3)$	GPR	0.25	0.24	0.21	0.24
$\epsilon(-2/-3)$	NN	0.25	0.22	0.22	0.24

FIG. 5: Random Forest Regression results: (a) parity plots, and (b) prediction uncertainty as a function of absolute prediction error.

they can be linearly interpreted. In this work, we use different possible choices for the kernel function, namely polynomial, radial basis function (RBF), and Laplacian. For hyperparameter optimization, we applied the grid search method to search a dense space for the best combination of kernel choice and different parameters in the kernel, separately for each output.

The prediction performances for the eight outputs are shown as parity plots in Fig. 6(a) and listed in Tables I and II. KRR shows a marked improvement in formation energy prediction and slight improvements in transition level predictions compared to RFR. The improvement is heavily owed to significant lowering of errors for impurities in the II-VI compounds. We

TABLE II: ML test set prediction RMSE values for formation energies.

Property	ML Method	II-VI Error (eV)	III-V Error (eV)	IV-IV Error (eV)	Total Error (eV)
ΔH (A-rich)	MLR	0.85	1.57	1.81	1.16
ΔH (A-rich)	Ridge	0.85	1.54	1.78	1.14
ΔH (A-rich)	LASSO	0.88	1.55	1.79	1.16
ΔH (A-rich)	Elastic Net	0.85	1.53	1.78	1.14
ΔH (A-rich)	RFR	1.05	1.03	1.20	1.07
ΔH (A-rich)	KRR	0.62	1.35	1.32	0.89
ΔH (A-rich)	GPR	0.59	1.33	1.71	0.96
ΔH (A-rich)	NN	0.62	1.30	1.40	0.89
ΔH (B-rich)	MLR	1.04	1.82	1.81	1.31
ΔH (B-rich)	Ridge	1.04	1.73	1.77	1.29
ΔH (B-rich)	LASSO	1.08	1.74	1.80	1.32
ΔH (B-rich)	Elastic Net	1.05	1.72	1.77	1.28
ΔH (B-rich)	RFR	1.09	1.25	1.52	1.18
ΔH (B-rich)	KRR	0.77	1.52	1.45	1.03
ΔH (B-rich)	GPR	0.82	1.52	1.70	1.11
ΔH (B-rich)	NN	0.81	1.34	1.44	1.01

find the KRR RMSE for ΔH (A-rich) to be 0.89 eV and for ΔH (B-rich) to be 1.03 eV, while the RMSE values for the 6 transition levels range between 0.25 and 0.35 eV. As shown in Fig. 6(b), the uncertainties on the KRR predictions range from 0 to 0.25 eV for the transition levels and 0 to 1 eV for the formation energies. Once again, a large concentration of points lie in the low uncertainty, low error region, with a few outliers existing in the opposite end of the spectrum.

Gaussian Process Regression

Another non-linear regression technique that employs the “kernel trick” is Gaussian process regression (GPR)^{80,81}. GPR uses the kernel and the observation to define a likelihood function on account of the covariance of a prior distribution over the target functions. The prior and likelihood function is assumed to have a Gaussian distribution. Based on Bayes’ theorem⁸², we get a predictive posterior distribution, from which we can attain a point prediction using its mean, and an uncertainty value using its variance. A major difference between GPR and KRR is that GPR can internally choose each kernel’s hyperparameters by applying gradient-ascent on the marginal likelihood function, while KRR requires a grid or random search using a loss function.

It can be seen from the GPR parity plots in Fig. 7(a) and from Tables I and II that the prediction RMSE values are very similar to those obtained with KRR. The formation energy errors are between 0.96 and 1.1 eV while the transition level errors range from 0.24 to 0.36 eV. It can also be seen from the training prediction errors listed in Tables ?? and ?? that there is a larger difference between training and test RMSE for both formation energies and transition levels than KRR. This

can be explained by the flexibility of the GPR models which likely causes overfitting when dealing with a small dataset and high dimensional features⁸³. The uncertainty vs absolute error plots in Fig. 7(b) show similar trends to KRR, with a majority of the points occupying the low-error, low-uncertainty region.

Neural Network Regression

Finally, we used neural networks (NN) to train regression models and compared the results with nonlinear regression models from RFR, KRR, and GPR. The Keras functional API model was used to build a deep feed forward NN to machine learn a multi-output regression⁸⁴. A sequential model trained to predict the six transition levels and two formation energies was found to be time consuming and lacked the ability to predict multiple outputs at once effectively. Further, a grid search employed to explore the number of hidden layers, number of neurons, learning rate, epochs, batch size, optimizers, and activation functions was found to be inefficient. Separate models were thus trained for each property using the SISSO generated descriptors, and scikit-optimize (skopt) was utilized for Bayesian hyperparameter optimization. To overcome an overfitting problem arising from minimizing only the test RMSE, the optimization function was revised to also include the difference in train and test RMSE.

Each NN architecture contains two to three dense neuron layers, through which the input is concatenated before returning the output through the final layer. The number of neurons in each dense layer varies with the input dimensions for each specific property or output. Kernel and activity regularizers were also integrated in each dense layer to prevent overfitting. The “relu” activation function was ultimately used for

FIG. 6: Kernel Ridge Regression results: (a) parity plots, and (b) prediction uncertainty as a function of absolute prediction error.

each dense layer, beating out sigmoid, softmax, softplus, tanh, and selu functions⁸⁵, while the Adam optimizer was selected over SGD, RMSEprop, Adadelata, and Adagrad⁸⁶. NN model training involved 10-repeated 5-fold cross-validation, where the mean and standard deviation of prediction of every data point were used as the predicted value and uncertainty value,

respectively. Parity plots for the best models thus obtained are presented in Fig. 8(a), while Fig. 8(b) shows the uncertainty vs absolute error plots.

It can be seen from Fig. 8 and Tables I and II that NN predictions for both transition levels and formation energies are similar to KRR and GPR. Transition level RMSE values are

FIG. 7: Gaussian Process Regression results: (a) parity plots, and (b) prediction uncertainty as a function of absolute prediction error.

seen to range from 0.24 to 0.33 eV while the formation energy RMSEs are between 0.9 and 1 eV. A comparison with training set predictions in Tables ?? and ?? further reveals that the gaps in test and training predictions from NN are similar to those from KRR, implying less overfitting as compared to GPR. A possible disadvantage of the NN models comes

from the larger uncertainty values seen in general compared to other methods, as visible from Fig. 8(b), while the absolute error values are similar to other methods. This is an effect of the stronger dependence of NN model prediction on the hyperparameter choice, leading to a larger standard deviation in prediction; this is expected to affect NN predictions over the

FIG. 8: Neural Network Regression results: (a) parity plots, and (b) prediction uncertainty as a function of absolute prediction error.

entire chemical space.

HIGH-THROUGHPUT SCREENING OF DOMINATING IMPURITIES

The detailed machine learning analysis presented in this work reveals that multiple nonlinear regression techniques can be

trained to make predictions of impurity transition levels and formation energies with errors that are within 10% of the range of values across the dataset. In Fig. 9(a), we present the test set prediction RMSE values of 8 different ML techniques used in this work, namely MLR, ridge, LASSO, elastic net, RFR, KRR, GPR and NN, for the six transition levels and two formation energies. The errors are plotted separately for the II-VI, III-V, and IV-IV points, as well as all the points taken together. It can be seen that for all the data types, the general trend is a reduction in RMSE upon going from linear to non-linear techniques. It is also seen that in general, the RFR performance is worse than KRR, GPR, and NN, while the latter three have similar formation energy errors with NN edging out the other two for most of the transition levels. From these results, one can expect NN, GPR, and KRR to yield similar results for the complete formation energy picture of all impurities as a function of charge, chemical potential, and Fermi level, which can be formulated using the predicted neutral state formation energies and all possible charge transitions.

We performed high-throughput prediction of the complete formation energies of the entire dataset of 12,474 impurities, using the best NN, GPR, KRR, and RFR models. The predicted impurity formation energies were then compared with the dominant native defect energetics for each compound, based on which screening is performed for (a) dominating impurities, i.e., impurities with lower energy than native defects which will change the equilibrium Fermi level of the semiconductor, and (b) low energy impurities (lower than native defects) with mid-gap energy levels. The screening performance of each ML model is determined by comparing the ML and DFT screening for the data points in the original DFT dataset. Given the expected DFT vs experiments and ML vs DFT errors, we relax the screening criteria by ± 0.2 eV for the DFT data and by ± 0.5 eV for the ML data. We thus calculated the number of true positives (TP, dominating/mid-gap from both DFT and ML), true negatives (TN), false positives (FP) and false negatives (FN) for each method. Based on these scores, the following metrics were defined:

- Accuracy = $(TP+TN)/(TP+TN+FP+FN)$
- Precision = $TP/(TP+FP)$
- Recall = $TP/(TP+FN)$

Figs. 9(b) and (c) show the accuracy, precision, and recall scores of each ML technique for screening of dominating impurities and low energy impurities with mid-gap levels, respectively. Results are plotted for the total dataset and for each semiconductor type, for both A-rich and B-rich conditions. The accuracies (in blue) of RFR, GPR, and KRR for all data types are seen to be greater than 95% for screening of dominating impurities in Fig. 9(b), while the precision (red) and recall (green) range from 80% to 95%. Interestingly, the accuracy, precision and recall scores of NN predictions are universally seen to lag behind the scores from RFR, GPR, and KRR. This surprising lack of predictive power of the NN models is attributed to their strong dependence on the hyperparameter choices, which is intimately linked with the exact nature of the training dataset. This leads to the higher uncer-

tainty values seen in Fig. 8(b) and likely overfitting which may not manifest in a limited test set, but over the entire set of 12,474 impurities, some predictions may be well off, resulting in lower accuracy, precision, and recall scores. The NN scores are better for screening of low energy impurities with mid-gap levels in Fig. 9(c), and all four techniques (NN, GPR, KRR and RF) show similar accuracy, precision and recall. Scores are seen to be lower for III-V data points than others, for reasons relating to the general imbalance in the dataset.

In Table III, we listed several impurities deemed to be dominating from both DFT and ML (GPR used as example here), along with their stable charge states, the corresponding dominating native defect, the type of shift induced in the equilibrium E_F , and whether mid-gap energy levels are created. For example, it can be seen that Ti at the Al site in AlAs creates a stable +1 charged donor type defect, and along with a -3 charged As vacancy acceptor, makes the conductivity more n-type and creates a transition level in the band gap. Similarly, a Be interstitial defect in Si induces a p-type shift in conductivity. Lists of dominating impurities with or without mid-gap energy levels were thus generated for all compounds. Finally, we plotted the complete charge and E_F dependent formation energies of selected impurities from both DFT and GPR for a few cases in Fig. 10. There is an impressive match between the DFT and GPR curves for most of the impurities, with charge states and transitions in general remaining consistent. A few impurities such as Bi_{Zn} in ZnS and In_i in AlAs are seen to show greater disparity between DFT and GPR, but qualitative trends remain the same. Also plotted for each case in Fig. 10 are the dominant native defects, and it can be seen that almost all impurities are correctly predicted to be dominating or not dominating from GPR compared to DFT, which implies a reliable qualitative screening even when actual predicted formation energies or transition levels are off.

PERSPECTIVE AND FUTURE WORK

The DFT+ML strategy presented in this work enables the quick prediction and screening of impurities in semiconductors, but is still limited by several factors. The primary concern is certainly the accuracy of the PBE functional which determines the reliability of the computational dataset and every subsequent step. Despite the impressive correspondence between measured and PBE computed defect levels, a generalization over all the semiconductor compounds and all types of impurities requires further caution. The use of advanced levels of theory, such as HSE06 and GW with and without SOC, may yet be necessary for future improvements of prediction models. However, ML models built on PBE data are still certainly useful for a number of reasons: (a) although quantitative predictions may be off, they provide qualitative screening of impurities likely to create low energy charged defects and/or consequential energy levels in the band gap, with an expected accuracy of $> 95\%$, and (b) PBE and ML-PBE estimates provide starting points for more advanced calculations, and can be used in a multi-fidelity learning framework wherein higher

FIG. 9: The performance of various ML models by semiconductor type, in terms of (a) prediction RMSE, and screening accuracy, precision and recall scores for (b) dominating impurities and (c) low energy impurities with mid-gap energy levels.

FIG. 10: A comparison of the complete charge and Fermi level dependent formation energy picture of selected impurities from DFT (solid lines) and GPR (dashed lines), presented for (a) CdTe at Cd-rich conditions, (b) ZnS at S-rich conditions, (c) AlAs at As-rich conditions, (d) GaP at Ga-rich conditions, (e) Si at Si-rich conditions, and (f) SiC at C-rich conditions. Dominant donor and acceptor type native defects are also pictured.

fidelity predictions are improved using lower fidelity data. We note here that although we consider mid-gap states that only arise from defect charge transitions, there are other internal transitions such as the d-d or f-f transitions of transition metals and lanthanides that could potentially further affect the absorption and emission characteristics of a semiconductor^{87,88}.

Going forward, a number of extensions and improvements will be made to this work, the first being the generation of higher accuracy DFT data and training multi-fidelity learning models. In reference 12, we showed that with a much smaller set of HSE06 data points, ML descriptors could be combined with models trained on larger quantities of PBE data to yield excellent predictions for Cd-chalcogenides; this will be extended to all group IV, III-V, and II-VI semiconductors. Various types of multi-fidelity learning models can be developed, from PBE-experiments (using the current dataset of 89 points supplemented with more data) to PBE-HSE to PBE-HSE-experiments, providing a potential pathway to bridging the DFT vs. experiment gap. Further, while all compounds were currently studied in the zinc blende structure, the DFT data and ML models will be extended to include defects in the wurtzite and rocksalt structures. The current ML framework can also be extended to semiconductor alloys in the same chemical spaces using the same type of descriptors, as was demonstrated for the limited example of Cd-chalcogenides¹².

The set of descriptors and ML methods used can also be expanded, for instance by including low accuracy unit cell defect calculations as used in references^{6,12}. Finally, tools can be created for on-demand prediction of the entire defect formation energy picture of any point defect or impurity in any compound, and a comparison of said defect with dominating native defects and other impurities.

CONCLUSIONS

In summary, we used a combination of DFT and ML to predict the charge-, Fermi level-, and chemical potential-dependent formation energy of any substitutional or interstitial impurity or point defect in zinc blende structures of group IV, III-V and II-VI semiconductors. A DFT dataset was created for the neutral state formation energies and various charge transition levels of upwards of 1000 possible impurities across 34 compounds, which formed about 10% of the entire semiconductor+impurity chemical space. ML models were built from the data by using descriptors that included properties of the compound, the defect site, and the impurity atoms, and applying algorithms ranging from linear regression techniques to nonlinear methods such as random forest and neural net-

TABLE III: Selected dominating impurities identified by both DFT and ML (GPR), at A-rich chemical potential conditions.

Semiconductor	Impurity	Shift in Eqm. E_F	Dominating Defects	Mid-gap Level?
CdS	In_{Cd}	n-type	In_{Cd} , $q = 1$ and V_{Cd} , $q = -2$	Y
CdS	I_S	n-type	I_S , $q = 1$ and V_{Cd} , $q = -3$	Y
CdS	Ti_i	p-type	Ti_i , $q = 2$ and V_S , $q = -1$	Y
CdSe	Cu_{Cd}	p-type	Cu_{Cd} , $q = -1$ and Cd_i , $q = 2$	Y
CdSe	F_i	p-type	F_i , $q = -1$ and V_{Se} , $q = 2$	N
CdSe	Ni_i	p-type	Ni_i , $q = -1$ and V_{Se} , $q = 2$	Y
CdTe	Bi_{Cd}	n-type	Bi_{Cd} , $q = 1$ and V_{Cd} , $q = -2$	Y
CdTe	As_{Te}	p-type	As_{Te} , $q = -1$ and V_{Te} , $q = 2$	Y
CdTe	Na_i	n-type	Na_i , $q = 1$ and V_{Cd} , $q = -2$	N
ZnS	Li_i	n-type	Li_i , $q = 1$ and V_{Zn} , $q = -2$	N
ZnS	Ti_i	n-type	Ti_i , $q = 1$ and V_{Zn} , $q = -2$	Y
ZnSe	Al_{Zn}	n-type	Al_{Zn} , $q = 1$ and V_{Zn} , $q = -2$	Y
ZnSe	Br_{Se}	n-type	Br_{Se} , $q = 1$ and Zn_{Se} , $q = -1$	Y
ZnTe	Cr_i	n-type	Cr_i , $q = 1$ and V_{Te} , $q = -2$	N
ZnTe	Mn_i	n-type	Mn_i , $q = 1$ and Zn_{Te} , $q = -2$	Y
AlN	Se_N	p-type	Se_N , $q = -1$ and V_N , $q = 1$	Y
AlP	Hf_{Al}	n-type	Hf_{Al} , $q = 1$ and Al_P , $q = -1$	Y
AlP	Cr_i	n-type	Cr_i , $q = 1$ and V_{Al} , $q = -2$	Y
AlAs	Ti_{Al}	n-type	Ti_{Al} , $q = 1$ and V_{As} , $q = -3$	Y
GaN	Tl_{Ga}	p-type	Tl_{Ga} , $q = -1$ and V_N , $q = 1$	Y
GaN	P_N	p-type	P_N , $q = -2$ and V_N , $q = 1$	Y
GaP	Ni_{Ga}	p-type	Ni_{Ga} , $q = -1$ and Ga_i , $q = 2$	Y
GaP	Li_i	n-type	Li_i , $q = 1$ and Ga_P , $q = -2$	Y
GaAs	Sc_i	n-type	Sc_i , $q = 3$ and Ga_{As} , $q = -2$	Y
GaSb	Al_{Ga}	n-type	Al_{Ga} , $q = 1$ and V_{Ga} , $q = -2$	Y
InN	Zr_i	n-type	Zr_i , $q = 2$ and V_N , $q = -1$	Y
InP	Cu_i	n-type	Cu_i , $q = 1$ and In_P , $q = -2$	Y
InAs	Ca_{In}	p-type	Ca_{In} , $q = -1$ and In_{As} , $q = 2$	N
Si	Ti_{Si}	p-type	Ti_{Si} , $q = -1$ and Si_i , $q = 2$	Y
Si	Be_i	n-type	Be_i , $q = 1$ and V_{Si} , $q = -3$	Y
SiC	V_{Si}	n-type	V_{Si} , $q = 1$ and V_C , $q = -2$	Y
SiC	Cr_i	p-type	Cr_i , $q = -1$ and V_C , $q = 1$	Y
SnC	As_{Sn}	n-type	As_{Sn} , $q = 1$ and V_C , $q = -2$	N
SnC	Cr_{Sn}	p-type	Cr_{Sn} , $q = -1$ and V_C , $q = 2$	N

works. For the 8 properties of interest (2 formation energies and 6 transition levels), KRR, GPR, and NN generally lead to similar performances, and the best models were deployed to predict all impurity properties in a high-throughput manner. Lists of dominating impurities, which can change the equilibrium conductivity of the compound as determined by native defects, were created using the ML predictions. The learning and design framework described in this work can be extended in terms of new semiconductors and mixed composition compounds, more involved descriptors and ML techniques, and more advanced levels of theory. The same design framework is also applicable to other semiconductor classes such as halide perovskites and I-III-VI semiconductors, and can lead to novel materials with improved optoelectronic properties for solar cells and related applications.

Data Availability

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

Code Availability

An initial version of a tool to facilitate the AI-driven prediction and screening of point defects and impurities in semiconductors can be found here: https://github.com/lmjacoby/ai_semiconductors.

ACKNOWLEDGMENTS

This work was performed, in part, at the Center for Nanoscale Materials, a U.S. Department of Energy Office of Science User Facility, and supported by the U.S. Department of Energy, Office of Science, under Contract No. DE-AC02-06CH11357. A.M.K., X.X. and L.J. contributed equally to this work. X.X., L.J. and R.B. acknowledge support from the Data Intensive Research Enabling Clean Technology (DIRECT) NSF National Research Traineeship. X.X., L.J., D.G. and S.D. acknowledge support from the UW Molecular Engineering Materials Center (DMR 1719797), an NSF Materials Research Science and Engineering Center. 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. 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. S.D. acknowledges funding from the U.S. Department of Energy, award number DE-EE0008556. A.M.K. acknowledges support from the School of Materials Engineering at Purdue University under account number F.10023800.05.002.

Competing Interests

The authors declare no competing financial or non-financial interests.

Author Contributions

M.K.Y.C. and A.M.K. conceived the idea. A.M.K. and M.K.Y.C. performed the DFT computations. X.X., L.J., R.B. and A.M.K. performed the machine learning analysis. All authors contributed to the discussion and writing of the manuscript.

Additional Information

Correspondence and requests for materials should be addressed to A.M.K. (email: amannodi@purdue.edu) or M.K.Y.C. (email: mchan@anl.gov).

REFERENCES

- Compositional engineering of multinary Cu-In-Zn-based semiconductor nanocrystals for efficient and solution-processed red-emitting quantum-dot light-emitting diodes. *Organic Electronics* **74**, 46–51 (2019).
- Oba, F. & Kumagai, Y. Design and exploration of semiconductors from first principles: A review of recent advances. *Applied Physics Express* **11**, 060101 (2018). URL <https://doi.org/10.7567/apex.11.060101>.
- Ning, C.-Z., Dou, L. & Yang, P. Bandgap engineering in semiconductor alloy nanomaterials with widely tunable compositions. *Nature Reviews Materials* **2** (2017). URL <https://doi.org/10.1038/natrevmats.2017.70>.
- Klug, M. T. *et al.* Tailoring metal halide perovskites through metal substitution: influence on photovoltaic and material properties. *Energy Environ. Sci.* **10**, 236–246 (2017). URL <http://dx.doi.org/10.1039/C6EE03201J>.
- Sampson, M. D., Park, J. S., Schaller, R. D., Chan, M. K. Y. & Martinson, A. B. F. Transition metal-substituted lead halide perovskite absorbers. *J. Mater. Chem. A* **5**, 3578–3588 (2017).
- Mannodi-Kanakkithodi, A. *et al.* Comprehensive computational study of partial lead substitution in methylammonium lead bromide. *Chemistry of Materials* **31**, 3599–3612 (2019).
- Queisser, H. J. & Haller, E. E. Defects in semiconductors: Some fatal, some vital. *Science* **281**, 945–950 (1998). URL <http://science.sciencemag.org/content/281/5379/945>. <http://science.sciencemag.org/content/281/5379/945.full.pdf>.
- Krasikov, D. & Sankin, I. Defect interactions and the role of complexes in the CdTe solar cell absorber. *J. Mater. Chem. A* **5**, 3503–3513 (2017). URL <http://dx.doi.org/10.1039/C6TA09155E>.
- Yin, W.-J., Shi, T. & Yan, Y. Unusual defect physics in CH₃NH₃PbI₃ perovskite solar cell absorber. *Applied Physics Letters* **104**, 063903 (2014).
- Park, J. S., Kim, S., Xie, Z. & Walsh, A. Point defect engineering in thin-film solar cells. *Nature Reviews Materials* **2058-8437** (2018).
- Mannodi-Kanakkithodi, A. *et al.* Machine-learned impurity level prediction for semiconductors: the example of Cd-based chalcogenides. *npj Computational Materials* **6** (2020).
- Mannodi-Kanakkithodi, A., Park, J.-S., Martinson, A. B. F. & Chan, M. K. Y. Defect energetics in pseudo-cubic mixed halide lead perovskites from first-principles. *The Journal of Physical Chemistry C* **124**, 16729–16738 (2020).
- Sivathanu, V., R. T. & Lenka, T. R. Modeling and performance optimization of two-terminal Cu₂ZnSnS₄-silicon tandem solar cells. *International Journal of Energy Research* **n/a**, 1–11 (2021).
- Liu, Z. *et al.* Computational functionality-driven design of semiconductors for optoelectronic applications. *InfoMat* **2**, 879–904 (2020).
- Wang, J., Chen, H., Wei, S.-H. & Yin, W.-J. Materials design of solar cell absorbers beyond perovskites and conventional semiconductors via combining tetrahedral and octahedral coordination. *Advanced Materials* **31**, 1806593 (2019).
- Nayak, P. K., Mahesh, S., Snaith, H. J. & Cahen, D. Photovoltaic solar cell technologies: analysing the state of the art. *Nature Reviews Materials* **4**, 269–285 (2019).
- Torres-Jaramillo, Santiago, Bernal-Correa, Roberto & Morales-Acevedo, Arturo. Improved design of InGaP/GaAs/InGaP tandem solar cells. *EPJ Photovoltaic* **12**, 1 (2021). URL <https://doi.org/10.1051/epjpv/2021001>.
- Energy-Band Structure: *Energy-Band Gaps*, chap. 6, 103–145 (John Wiley Sons, Ltd, 2005). URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/0470090340.ch6>. <https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470090340.ch6>.
- Bouarissa, N., Algarni, H., Ajmal Khan, M., Al-Hagan, O. & Alhuwaymel, T. Collective effects and optical characteristics of CdSe_xTe_{1-x}. *Optik* **203**, 163952 (2020). URL <https://www.sciencedirect.com/science/article/pii/S0030402619318509>.
- Dumre, B. *et al.* Improved optoelectronic properties in CdSe_xTe_{1-x} through controlled composition and short-range order. *Solar Energy* **194**, 742–750 (2019). URL <https://www.sciencedirect.com/science/article/pii/S0038092X19310874>.
- Ban, D. *et al.* Optimized GaAs/GaAs light-emitting diodes and high efficiency wafer-fused optical up-conversion devices. *Journal of Applied Physics* **96**, 5243–5248 (2004).
- Wang, G., Yi, X., Zhan, T. & Huang, Y. *The AlGaInP/AlGaAs Material System and Red/Yellow LED*, 171–202 (Springer International Publishing, Cham, 2019). URL https://doi.org/10.1007/978-3-319-99211-2_5.

- ²³Heo, S. *et al.* Deep level trapped defect analysis in $\text{CH}_3\text{NH}_3\text{PbI}_3$ perovskite solar cells by deep level transient spectroscopy. *Energy Environ. Sci.* **10**, 1128–1133 (2017). URL <http://dx.doi.org/10.1039/C7EE00303J>.
- ²⁴Rosenberg, 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. *Journal of Applied Physics* **122**, 145701 (2017).
- ²⁵Schultz, P. A. Theory of defect levels and the “band gap problem” in silicon. *Phys. Rev. Lett.* **96**, 246401 (2006). URL <https://link.aps.org/doi/10.1103/PhysRevLett.96.246401>.
- ²⁶Computational data-driven materials discovery. *Trends in Chemistry* **3**, 79–82 (2021). URL <https://www.sciencedirect.com/science/article/pii/S2589597420303178>. Special Issue: Machine Learning for Molecules and Materials.
- ²⁷Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials* **3**, 54 (2017). URL <https://doi.org/10.1038/s41524-017-0056-5>.
- ²⁸Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* **5**, 83 (2019). URL <https://doi.org/10.1038/s41524-019-0221-0>.
- ²⁹Kim, C., Chandrasekaran, A., Huan, T. D., Das, D. & Ramprasad, R. Polymer genome: A data-powered polymer informatics platform for property predictions. *The Journal of Physical Chemistry C* **122**, 17575–17585 (2018).
- ³⁰Ward, L. *et al.* Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **152**, 60–69 (2018). URL <https://www.sciencedirect.com/science/article/pii/S0927025618303252>.
- ³¹Agrawal, A. & Choudhary, A. An online tool for predicting fatigue strength of steel alloys based on ensemble data mining. *International Journal of Fatigue* **113**, 389–400 (2018). URL <https://www.sciencedirect.com/science/article/pii/S0142112318301476>.
- ³²Doan, H. A. *et al.* Quantum chemistry-informed active learning to accelerate the design and discovery of sustainable energy storage materials. *Chemistry of Materials* **32**, 6338–6346 (2020).
- ³³Mannodi-Kanakkithodi, A. *et al.* Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond. *Materials Today* **21**, 785–796 (2018). URL <https://www.sciencedirect.com/science/article/pii/S1369702117307344>.
- ³⁴Khmaissia, F. *et al.* Accelerating band gap prediction for solar materials using feature selection and regression techniques. *Computational Materials Science* **147**, 304–315 (2018). URL <https://www.sciencedirect.com/science/article/pii/S0927025618300958>.
- ³⁵Feng, H.-J., Wu, K. & Deng, Z.-Y. Predicting inorganic photovoltaic materials with efficiencies >26% via structure-relevant machine learning and density functional calculations. *Cell Reports Physical Science* **1**, 100179 (2020). URL <https://www.sciencedirect.com/science/article/pii/S2666386420301909>.
- ³⁶Iwasaki, Y. *et al.* Machine-learning guided discovery of a new thermoelectric material. *Scientific Reports* **9**, 2751 (2019).
- ³⁷Meredig, B. *et al.* Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, 094104 (2014). URL <https://link.aps.org/doi/10.1103/PhysRevB.89.094104>.
- ³⁸Vasudevan, R. K. *et al.* Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics. *MRS Communications* **9**, 821–838 (2019).
- ³⁹Chan, M. & Ceder, G. Efficient band gap prediction for solids. *Physical Review Letters* **105**, 196403 (2010).
- ⁴⁰Heyd, 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. *The Journal of Chemical Physics* **123**, 174101 (2005).
- ⁴¹Aryasetiawan, F. & Gunnarsson, O. The GW method. *Reports on Progress in Physics* **61**, 237–312 (1998).
- ⁴²Grimmeiss, H. G. Deep level impurities in semiconductors. *Annual Review of Materials Science* **7**, 341–376 (1977).
- ⁴³Chen, J. W. & Milnes, A. G. Energy levels in silicon. *Annual Review of Materials Science* **10**, 157–228 (1980).
- ⁴⁴Dow, J. D., Ren, S. Y. & Shen, J. *Deep Impurity Levels in Semiconductors, Semiconductor Alloys, and Superlattices*, 175–187 (Springer US, Boston, MA, 1988). URL https://doi.org/10.1007/978-1-4684-5553-3_14.
- ⁴⁵Itoh, H. *et al.* Intrinsic defects in cubic silicon carbide. *physica status solidi (a)* **162**, 173–198 (1997).
- ⁴⁶Schöler, M., Lederer, M. W., Schuh, P. & Wellmann, P. J. Intentional incorporation and tailoring of point defects during sublimation growth of cubic silicon carbide by variation of process parameters. *physica status solidi (b)* **257**, 1900286 (2020).
- ⁴⁷Lebedev, A. A. Deep level centers in silicon carbide: A review. *Semiconductors* **33**, 107–130 (1999).
- ⁴⁸Zhou, P., Spencer, M. G., Harris, G. L. & Fekade, K. Observation of deep levels in cubic silicon carbide. *Applied Physics Letters* **50**, 1384–1385 (1987).
- ⁴⁹Nagesh, V., Farmer, J. W., Davis, R. F. & Kong, H. S. Defects in cubic silicon. *Radiation Effects and Defects in Solids* **112**, 77–84 (1990).
- ⁵⁰Itoh, H. *et al.* Photoluminescence of radiation induced defects in 3c-sic epitaxially grown on si. *Journal of Applied Physics* **77**, 837–842 (1995).
- ⁵¹Swaminathan, V. Defects in GaAs. *Bulletin of Materials Science* **4**, 403–442 (1982).
- ⁵²Lindström, A., Mirbt, S., Sanyal, B. & Klintonberg, M. High resistivity in undoped CdTe: carrier compensation of Te antisites and Cd vacancies. *Journal of Physics D: Applied Physics* **49**, 035101 (2015).
- ⁵³Ablekim, T. *et al.* Self-compensation in arsenic doping of CdTe. *Scientific Reports* **7**, 4563 (2017).
- ⁵⁴Komin, 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).
- ⁵⁵Ayoub, M. *et al.* Annealing effects on defect levels of CdTe:Cl materials and the uniformity of the electrical properties. *IEEE Transactions on Nuclear Science* **50**, 229–237 (2003).
- ⁵⁶Jantsch, W. & Hendorfer, G. Characterization of deep levels in CdTe by photo-epr and related techniques. *Journal of Crystal Growth* **101**, 404–413 (1990). URL <http://www.sciencedirect.com/science/article/pii/002202489091004A>.
- ⁵⁷Kraft, C. *et al.* Phosphorus implanted cadmium telluride solar cells. *Thin Solid Films* **519**, 7153–7155 (2011). URL <http://www.sciencedirect.com/science/article/pii/S0040609011004573>. Proceedings of the EMRS 2010 Spring Meeting Symposium M: Thin Film Chalcogenide Photovoltaic Materials.
- ⁵⁸Kresse, G. & Hafner, J. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B* **49**, 14251–14269 (1994). URL <https://link.aps.org/doi/10.1103/PhysRevB.49.14251>.
- ⁵⁹Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996). URL <https://link.aps.org/doi/10.1103/PhysRevB.54.11169>.
- ⁶⁰Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996). URL <https://link.aps.org/doi/10.1103/PhysRevLett.77.3865>.
- ⁶¹Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994). URL <https://link.aps.org/doi/10.1103/PhysRevB.50.17953>.
- ⁶²Freyssoldt, 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). URL <https://link.aps.org/doi/10.1103/PhysRevLett.102.016402>.
- ⁶³Freyssoldt, C. *et al.* First-principles calculations for point defects in solids. *Rev. Mod. Phys.* **86**, 253–305 (2014). URL <https://link.aps.org/doi/10.1103/RevModPhys.86.253>.
- ⁶⁴Schultz, P. A. First principles predictions of intrinsic defects in aluminum arsenide, AlAs. *MRS Proceedings* **1370**, mrs11–1370-yy03–04 (2011).
- ⁶⁵Olivetti, E. A. *et al.* Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews* **7**, 041317 (2020).
- ⁶⁶Kononova, O. *et al.* Opportunities and challenges of text mining in materials research. *iScience* **24**, 102155 (2021). URL <https://www.>

- sciencedirect.com/science/article/pii/S2589004221001231.
- ⁶⁷Sun, R., Chan, M. K. Y., Kang, S. & Ceder, G. Intrinsic stoichiometry and oxygen-induced *p*-type conductivity of pyrite FeS_2 . *Phys. Rev. B* **84**, 035212 (2011). URL <https://link.aps.org/doi/10.1103/PhysRevB.84.035212>.
- ⁶⁸FELton, D. C., Boukouvalas, Z., Butrico, M. S., Fuge, M. D. & Chung, P. W. Applying machine learning techniques to predict the properties of energetic materials. *Scientific Reports* **8**, 9059 (2018).
- ⁶⁹Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M. & Ghiringhelli, L. M. Sisso: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Materials* **2**, 083802 (2018). URL <https://link.aps.org/doi/10.1103/PhysRevMaterials.2.083802>.
- ⁷⁰Fan, J. & Lv, J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849–911 (2008). URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2008.00674.x>. <https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2008.00674.x>.
- ⁷¹Fan, J., Samworth, R. & Wu, Y. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of machine learning research* **10**, 2013–2038 (2009).
- ⁷²Kramer, O. *K-Nearest Neighbors*, 13–23 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013). URL https://doi.org/10.1007/978-3-642-38652-7_2.
- ⁷³George, A. & Vidyapeetham, A. Anomaly detection based on machine learning: dimensionality reduction using pca and classification using svm. *International Journal of Computer Applications* **47**, 5–8 (2012).
- ⁷⁴Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K. & Chang, L. A novel anomaly detection scheme based on principal component classifier. In *Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM'03)*, 172–179 (2003).
- ⁷⁵Varoquaux, G. *et al.* Scikit-learn: Machine learning without learning the machinery. *GetMobile: Mobile Comp. and Comm.* **19**, 29–33 (2015). URL <https://doi.org/10.1145/2786984.2786995>.
- ⁷⁶Friedman, J., Hastie, T. & Tibshirani, R. *The elements of statistical learning*, vol. 1 (Springer series in statistics New York, 2001).
- ⁷⁷Ng, A. Y. Feature selection, ℓ_1 vs. ℓ_2 regularization, and rotational invariance. In *Proceedings of the Twenty-First International Conference on Machine Learning, ICML '04*, 78 (Association for Computing Machinery, New York, NY, USA, 2004). URL <https://doi.org/10.1145/1015330.1015435>.
- ⁷⁸Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).
- ⁷⁹Bergstra, J., Yamins, D. & Cox, D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on Machine Learning*, vol. 28 (JMLR, Atlanta, Georgia, USA, 2013).
- ⁸⁰Williams, C. K. & Rasmussen, C. E. *Gaussian processes for machine learning*, vol. 2 (MIT press Cambridge, MA, 2006).
- ⁸¹SEEGER, M. Gaussian processes for machine learning. *International Journal of Neural Systems* **11**, 69–106 (2004).
- ⁸²Puga, J. L., Krzywinski, M. & Altman, N. Bayes' theorem **12**, 277–278 (2015).
- ⁸³Mohammed, R. O. & Cawley, G. C. Over-fitting in model selection with gaussian process regression. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, 192–205 (Springer, 2017).
- ⁸⁴Bisong, E. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* (2019).
- ⁸⁵Ramachandran, P., Zoph, B. & Le, Q. V. Searching for activation functions (2018). URL <https://openreview.net/forum?id=SkBYyZRZ>.
- ⁸⁶Choi, D. *et al.* On empirical comparisons of optimizers for deep learning (2020). URL <https://openreview.net/forum?id=HygrAR4tPS>.
- ⁸⁷Sugano, S., Tanabe, Y. & Kamimura, H. *Multiplets of Transition-Metal Ions in Crystals* (Academic Press, 1970).
- ⁸⁸Fazzio, A., Caldas, M. J. & Zunger, A. Many-electron multiplet effects in the spectra of 3d impurities in heteropolar semiconductors. *Phys. Rev. B* **30**, 3430–3455 (1984). URL <https://link.aps.org/doi/10.1103/PhysRevB.30.3430>.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SI.pdf](#)