Article

Patterns

Universal machine learning framework for defect predictions in zinc blende semiconductors

Highlights

- Large computational dataset of defect properties in semiconductors is developed
- Regression algorithms are used to train predictive models for defect properties
- Best models are used for high-throughput prediction and screening
- Lists of low energy "dominating" impurities are generated

Authors

Arun Mannodi-Kanakkithodi, Xiaofeng Xiang, Laura Jacoby, Robert Biegaj, Scott T. Dunham, Daniel R. Gamelin, Maria K.Y. Chan

Correspondence

amannodi@purdue.edu (A.M.-K.), mchan@anl.gov (M.K.Y.C.)

In brief

A novel and universal ML framework is developed to predict charge-, Fermi level-, and chemical potential-dependent formation energies of point defects in zinc blende semiconductors such as CdTe, GaAs, SiC, and so on. We lay out in detail the methodology for data generation using high-throughput DFT simulations, training of ML models via feature selection, hyperparameter optimization, cross-validation, and prediction and screening for a dataset of more than 12,000 point defects across 34 compounds.

Mannodi-Kanakkithodi et al., 2022, Patterns *3*, 100450
March 11, 2022
https://doi.org/10.1016/j.patter.2022.100450

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Universal machine learning framework for defect predictions in zinc blende semiconductors

Arun Mannodi-Kanakkithodi,[1,2,7,8,*] Xiaofeng Xiang,[3,7] Laura Jacoby,[4,7] Robert Biegaj,[5] Scott T. Dunham,[6] Daniel R. Gamelin,[4] and Maria K.Y. Chan[1,*]

[1]Center for Nanoscale Materials, Argonne National Laboratory, Argonne, IL 60439, USA
[2]School of Materials Engineering, Purdue University, West Lafayette, IN 47907, USA
[3]Molecular Engineering & Sciences Institute, University of Washington, Seattle, WA 98195, USA
[4]Department of Chemistry, University of Washington, Seattle, WA 98195, USA
[5]Materials Science & Engineering, University of Washington, Seattle, WA 98195, USA
[6]Department of Electrical and Computer Engineering, University of Washington, Seattle, WA 98195, USA
[7]These authors contributed equally
[8]Lead contact
*Correspondence: amannodi@purdue.edu (A.M.-K.), mchan@anl.gov (M.K.Y.C.)
https://doi.org/10.1016/j.patter.2022.100450

THE BIGGER PICTURE Our article introduces a universal predictive framework for point defect formation energies and charge transition levels in a wide chemical space of zinc blende semiconductors and possible impurity atoms selected from across the periodic table. This framework was developed by leveraging high-throughput quantum mechanical simulations benchmarked using some experimental data from the literature, as well as machine learning (ML)-based regressions techniques that map unique materials descriptors to computed defect properties and yield optimized and generalizable models. The power and utility of these models is revealed through quick predictions for thousands of new defects and screening of low-energy impurities, which may tune the equilibrium conductivity in the semiconductor. This work presents, to our knowledge, the largest density functional theory (DFT) dataset of defect properties in semiconductors and the largest DFT+ML-based screening of point defects in semiconductors to date.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

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 groups 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 approximately 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 neutral state formation energies and charge transition levels of impurities. Optimized kernel ridge, Gaussian process, random forest, and neural network regression models are applied to screen impurities with lower formation energy than dominant native defects in all compounds.

INTRODUCTION

Compositional manipulation of semiconductors is one of the primary methods used 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 site. 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] A complete understanding of a

semiconductor's optoelectronic behavior requires estimating the formation energies of point defects, whether accidental or intentionally introduced.[3,11,12]

While approximately 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.[5,13–17] 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 properties.[18] 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 light-emitting diodes.[21,22]

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,[4,5,12,18] 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 paramount.[12] 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), the nature of equilibrium conductivity, and the 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 and cathodoluminescence,[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, the 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).[3,8,11,12,25] When an appropriate level of theory is applied, the DFT-computed defect charge transition levels have been seen to match well with measured levels and have helped to 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 calculations,[12] 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.[12,26–28] 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 ML. 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] Indeed, ML has been instrumental in accelerating the prediction of properties related to point defects and dopants in materials. This includes predicting vacancy formation and substitutional energies of oxides using regression algorithms applied on DFT data,[39–42] ML formation energies, transition levels, and the migration energies of point defects in known semiconductors and alloys,[43,44] predicting the dopability of semiconductors,[45] and improving high-fidelity predictions of point defect properties using previously unknown correlations.[46] Recent work from our group involved performing high-throughput DFT computations to study the formation energies and charge transition levels of impurities in halide perovskites[3] and Cd-chalcogenides,[12] 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 semiconductor and impurity chemical spaces as well as ML techniques can pave the path toward 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-chalcogenides,[12] 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, and GeC), III–V (e.g., BSb, GaAs, and InP), and II–VI (e.g., ZnSe and CdS) 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 six types of charge transition levels for any possible impurity. We used sure independence screening and sparsifying operator (SISSO) for feature selection and the 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 ML 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 mean 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,

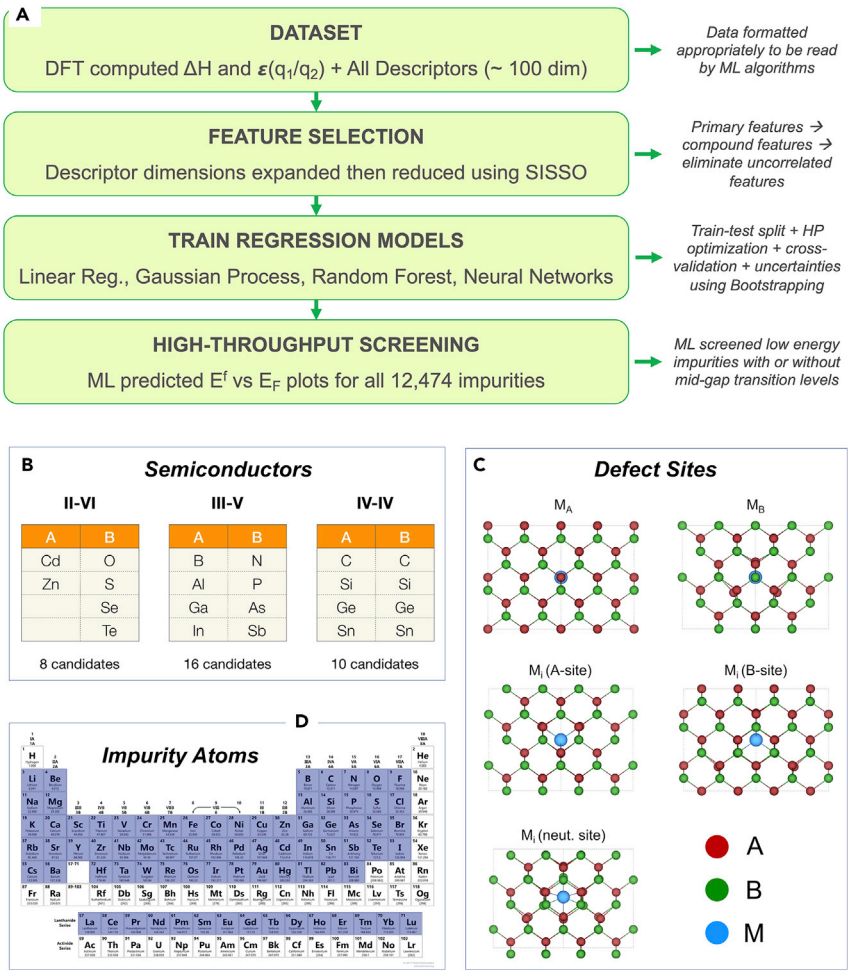

Figure 1. Outline and chemical space

(A–D) (A) The DFT-ML workflow followed in this work, and the semiconductor-impurity chemical space in terms of (B) the cation and anion choices for group IV, II–VI, and III–V compounds, (C) types of defect sites, and (D) impurity atoms selected from across the periodic table.

only approximately 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 workflow adopted here is laid out in Figure 1A, highlighting data generation, feature selection, regression model development, and high-throughput screening. The DFT data and ML codes generated through this work are made available on Github.

RESULTS

Semiconductor and impurity chemical space

The chemical space considered in this work has been pictured in Figure 1 in terms of the semiconductor compounds (b), possible defect sites (c), and impurity atoms (d). 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 group II–VI compounds (CdO, CdS, CdSe, CdTe, ZnO, ZnS, ZnSe, and ZnTe), 16 group 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 S1, along with corresponding experimental measurements collected from the literature. It can be seen that, although the cubic lattice constants are reasonably accurate, the standard generalized gradient approximation-Perdew-Burke-Ernzerhof (GGA-PBE) functional underestimates the band gap, as has been well demonstrated in the past.[47–49] 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. Figure 1 also shows the defect sites considered in this work, namely, the A and B sites and three types of interstitial sites: the A site interstitial (with 4 neighboring A atoms), the B site interstitial (with 4 neighboring B atoms), and the 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, and Sn), 3 defect sites are considered (A site, A site interstitial, and neutral site interstitial). For a few defects, we also tested other possible interstitial sites in the ZB structure, such as anion/cation-split sites (as described in the literature[7]), and generally found one of the three chosen interstitial sites to be lower in energy. 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 Figure 1. The total number of possible impurities in this chemical space can thus be estimated as: $77 \times 5 \times 30 + 77 \times 3 \times 4 = 12{,}474$. Out of these 12,474 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. The 10% of data points chosen for computations constituted a desirable diversity in semiconductor type, element type, and defect site type; in other words, while the actual data points were selected at random (and added to prior data[12]),

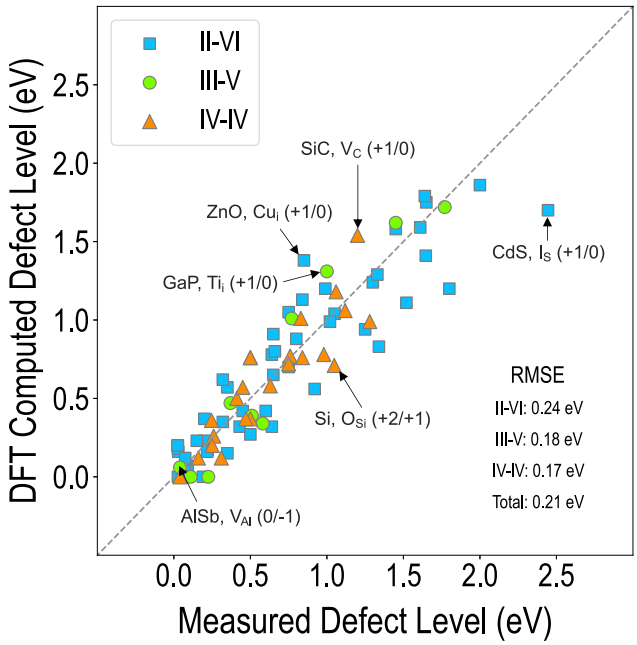

Figure 2. Comparison of DFT-computed defect levels with experimentally measured levels

(Obtained from publications[51–66]). Measured versus DFT RMSE values are also shown for different semiconductor types and for the combined set of points. A few defect levels have been labeled.

we ensured that every compound (out of 34), element (out of 77), and defect site (substitutional or interstitial) is roughly equally represented, with the exception of CdTe, which is heavily represented.

Defect properties: Benchmarking DFT and native defect energy picture

The methodology to compute defect formation energy (E^f) from DFT as a function of charge (q), chemical potential (μ) conditions, and Fermi level (E_F) is described in the Experimental procedures section. Ultimately, for all native defects and impurities, we calculate neutral state formation energies at two extreme μ conditions, namely A-rich and B-rich, and six types of defect charge 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 before 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 hybrid perovskites[3,10] and even group IV, III–V, and II–VI semiconductors.[11,25,46,50] This contrast can be attributed to total energy differences in DFT being more accurate than using Kohn-Sham energy levels to estimate band edges[11,25,50] or band gaps,[48] or in the case of hybrid perovskites, a fortuitous cancellation of errors that leads to PBE being as accurate as HSE06+SOC.[3,10] It has been reported in past work that defect levels computed from semilocal GGA-PBE for well-known ZB semiconductors such as Si and GaAs can span the physical band gap of the material[25,50]; that is, the defect transition levels calculated from PBE correspond well with experimental values up to the experimental band gap, even for transitions that are further from the VBM than the PBE-calculated band gap. 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[51–66] (in a brute force manner, something we hope to replace with more efficient and comprehensive natural language processing-based searches in the future[67,68]) and collected measured energy levels for 84 defects across the 34 compounds, adding to the set of 15 points collected for CdTe in ref[12] As presented in Figure 2 and Table S2, the PBE predicted defect levels are highly correlated with experimental measurements, with a correlation coefficient (R^2) of 0.85. We find that the RMSE of PBE predictions compared with experiments is less than 0.2 eV for III–V and IV–IV compounds, and 0.24 eV for II–VI compounds, resulting in a total RMSE of approximately 0.21 eV. This is an acceptable level of accuracy that is similar to what we found in earlier work,[12]and is within the recognized accuracy limit of DFT electronic levels; a similar ML versus 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 ML 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 Figure 3, 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 Figure 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, the 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 conditions[69]) 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 the V_{Zn}, meaning they cannot compensate for native defects. Similarly, the 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 Tl_{As} and C_i are both higher energy defects. In SiC at Si-rich conditions, the V_C and C_{Si} are the lowest energy donor and acceptor type defects and lead to intrinsic conductivity. Also marked in Figures 3A–3C are some neutral state

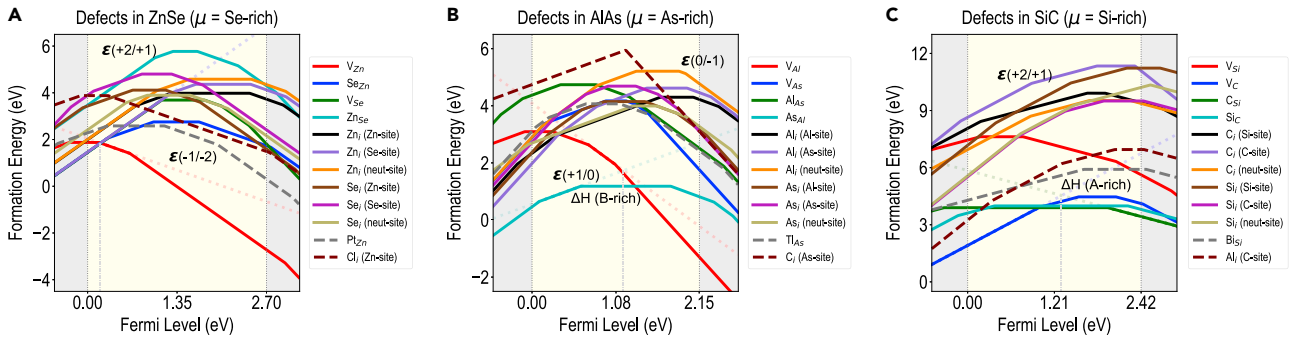

Figure 3. Charge- and Fermi-level-dependent formation energy picture

(A–C) 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 = 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.

formation energies, ΔH, and some charge transition levels, $\varepsilon(q_1/q_2)$. In Tables S3 and S4, we list the dominating acceptor and donor type defects, and the equilibrium E_F, E_F^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 on the dataset of impurity properties in Cd-chalcogenides from our previous work[12] 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 the formation energy as a function of the charge, Fermi level, and chemical potential, as shown in Figure 3. We explicitly considered two types of predictable properties that 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 ($\varepsilon(+3/+2)$, $\varepsilon(+2/+1)$, $\varepsilon(+1/0)$, $\varepsilon(0/-1)$, $\varepsilon(-1/-2)$ and $\varepsilon(-2/-3)$). 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 $\varepsilon(q_1/q_2)$ values for all six charge transition types. We then trained eight separate ML models for ΔH (A-rich), ΔH (B-rich), and transition levels from $\varepsilon(+3/+2)$ to $\varepsilon(-2/-3)$.

In Figure S1, 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 one-half of the data points belong to III–V semiconductors and one-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 approximately 40% substitutional impurities and 60% interstitial impurities, with the former being equally divided between A site and B site substitutions, and the latter divided equally between the three types of interstitial sites; all the defect sites are pictured in an example ZB supercell in Figure 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 Figure 4. The transition levels are plotted two at a time against each other in Figures 4A–4C, 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 Figure 4D, we plotted ΔH (A rich) versus ΔH (B rich), which shows values that range from approximately −5 eV to approximately 20 eV. For the group IV compounds C, Si, Ge, and Sn, the 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 and impurity chemical space used in this work, we can uniquely identify every data point using the identity of the semiconductor, the identity of

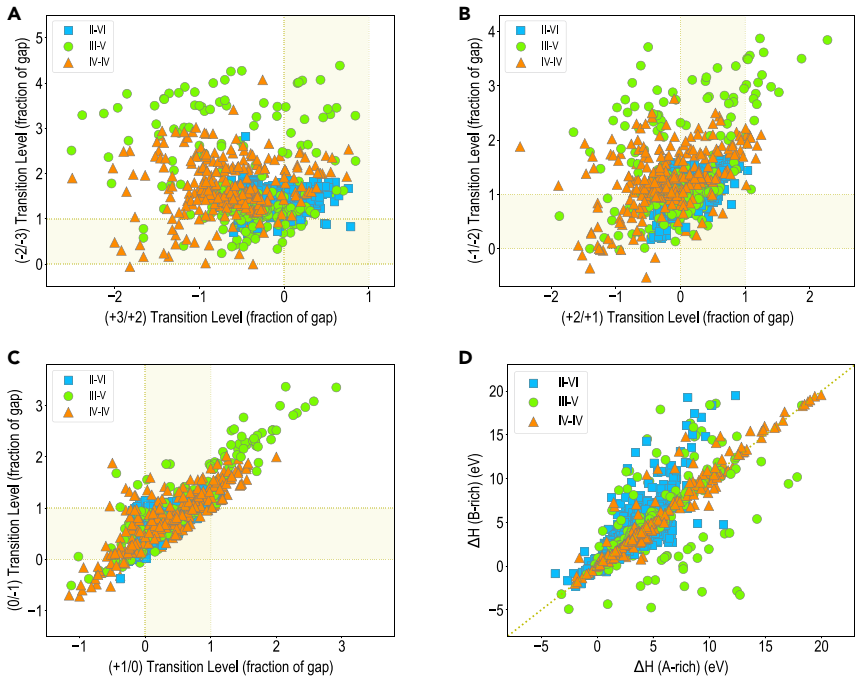

Figure 4. Visualization of DFT data

(A–D) (A–C) transition levels (+3/+2) to (−2/−3), and (D) neutral state formation energies at A-rich and B-rich chemical potential conditions, plotted for different semiconductor types.

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, namely, the formation energy, lattice constant, band gap, and the electronic and ionic dielectric constants; this leads to five dimensions.
2. $Elem_{prop}$: Tabulated elemental properties of the impurity M as well as species A and B, such as ionic radius, ionization energy, electronegativity, and so on; 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 definition[70]; this leads to eight dimensions.

The complete list of descriptors can be found on the x axis of Figure S2 as well as Table S5, which show the Pearson coefficient of linear correlation ($|r|$) between the properties of interest, ΔH and $\varepsilon(q_1/q_2)$, 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 SISSO[71] 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 approximately 10^2 to approximately 10^5 after two iterations. Next, sure independence screening[72] 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 operation[73] 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. KNN is a method commonly applied to identify the outliers in a dataset based on a feature space.[74] 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 the principal component analysis (PCA),[75] which decreases 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 that 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).[76] 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 used as the standard training sets.

Training regression models

In the following sections, we 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 the training-test split, cross-validation, hyperparameter optimization, and error evaluation was performed. A separate model was trained for each of the eight outputs, namely the six types of transition levels and two 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 to reduce the reported bias and variance, and is important for avoiding overfitting. Various important hyperparameters were optimized for

Table 1. 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)
ε(+3/+2)	MLR	0.35	0.37	0.34	0.35
ε(+3/+2)	Ridge	0.35	0.35	0.32	0.34
ε(+3/+2)	LASSO	0.36	0.36	0.32	0.35
ε(+3/+2)	Elastic net	0.35	0.35	0.32	0.34
ε(+3/+2)	RFR	0.36	0.31*	0.35	0.34
ε(+3/+2)	KRR	0.33	0.37	0.31	0.33
ε(+3/+2)	GPR	0.32	0.36	0.32	0.33
ε(+3/+2)	NN*	0.29*	0.36	0.29*	0.31*
ε(+2/+1)	MLR	0.42	0.46	0.46	0.44
ε(+2/+1)	Ridge	0.42	0.43	0.45	0.43
ε(+2/+1)	LASSO	0.43	0.44	0.45	0.44
ε(+2/+1)	Elastic net	0.42	0.43	0.45	0.43
ε(+2/+1)	RFR	0.39	0.36	0.40	0.38
ε(+2/+1)	KRR	0.33	0.38	0.40	0.36
ε(+2/+1)	GPR	0.32	0.38	0.41	0.36
ε(+2/+1)	NN*	0.29*	0.35*	0.38*	0.33*
ε(+1/0)	MLR	0.40	0.39	0.43	0.40
ε(+1/0)	Ridge	0.40	0.38	0.42	0.40
ε(+1/0)	LASSO	0.41	0.39	0.43	0.41
ε(+1/0)	Elastic net	0.40	0.38	0.42	0.40
ε(+1/0)	RFR	0.38	0.36	0.39	0.38
ε(+1/0)	KRR	0.31	0.34	0.38	0.33
ε(+1/0)	GPR*	0.29*	0.32	0.38	0.32*
ε(+1/0)	NN	0.29	0.31*	0.37*	0.32
ε(0/−1)	MLR	0.37	0.42	0.34	0.38
ε(0/−1)	Ridge	0.37	0.40	0.34	0.37
ε(0/−1)	LASSO	0.37	0.40	0.34	0.37
ε(0/−1)	Elastic net	0.37	0.40	0.34	0.37
ε(0/−1)	RFR	0.37	0.33	0.35	0.35
ε(0/−1)	KRR	0.32	0.36	0.32	0.33
ε(0/−1)	GPR	0.31	0.34	0.32	0.32
ε(0/−1)	NN*	0.28*	0.33*	0.31*	0.30*
ε(−1/−2)	MLR	0.33	0.38	0.30	0.33
ε(−1/−2)	Ridge	0.32	0.37	0.29	0.32
ε(−1/−2)	LASSO	0.32	0.37	0.29	0.33
ε(−1/−2)	Elastic net	0.32	0.37	0.29	0.33
ε(−1/−2)	RFR	0.34	0.35	0.27	0.33
ε(−1/−2)	KRR	0.29	0.32	0.27*	0.29
ε(−1/−2)	GPR	0.29	0.31	0.28	0.29
ε(−1/−2)	NN*	0.26*	0.29*	0.28	0.27*
ε(−2/−3)	MLR	0.27	0.26	0.22	0.26
ε(−2/−3)	Ridge	0.27	0.26	0.22	0.25
ε(−2/−3)	LASSO	0.27	0.26	0.22	0.25
ε(−2/−3)	Elastic net	0.27	0.26	0.22	0.25
ε(−2/−3)	RFR	0.24*	0.28	0.27	0.25
ε(−2/−3)	KRR	0.26	0.24	0.21	0.24
ε(−2/−3)	GPR	0.25	0.24	0.21	0.24
ε(−2/−3)	NN*	0.25	0.22*	0.22*	0.24*

*The gene in closest proximity to the cytokine QTL SNPs.

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, and so on. All regression models were trained using functions in the Python ML library scikit-learn.[77]

The metric for evaluating model performance was chosen to be the prediction RMSE. Each of the five 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; 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. 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 versus 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 S6 and S7, and the test prediction RMSE values listed in Tables 1 and 2, divided in terms of property, ML technique used, and type of data point. The set of best hyperparameter values obtained for each technique and each property are included as part of the Supplementary materials, as are alternative prediction performance metrics, mean absolute errors, and R_2 scores.

Linear regression

Figure S2 shows the Pearson coefficient of linear correlation between various (primary) descriptor dimensions and the properties of interest. We see that many of the features are 50%–70% correlated with the properties, showing a certain degree of linear relationship. We further plotted the correlation coefficients for the 10 best SISSO-based compound features (essentially, complex functions of combinations of original features) in Figure S3. The highest correlated features reveal the specific descriptors or combinations thereof that could best predict the defect formation energy and charge transition levels. We notice that atomic radii and ionization energy differences (between defect atom M and A/B atoms) are most important for $\varepsilon(+3/+2)$, while valence and electronegativity differences dominate for $\varepsilon(+2/+1)$; coefficients for both remain small, at approximately 0.35. The highest correlations are between 0.5 and 0.55 for $\varepsilon(+1/0)$ and $\varepsilon(0/-1)$, with ionization energy and atomic radii differences dominating in both. $\varepsilon(-1/-2)$ and $\varepsilon(-2/-3)$ show even higher correlations of between 0.7 and 0.75, with descriptors such as the Mendeleev number, covalent radii, and ICSD volume of M/A/B atoms being most important. Finally, we find correlation maximums of approximately 0.45 for ΔH (A rich) and ΔH (B rich), determined primarily by the semiconductor lattice constant, ionization potential, boiling point, heat of fusion/vaporization, and specific heat capacity. Overall, these correlations reveal that the relative electronegativities, ionization energies, and radii of elements are important in placing defect energy

Table 2. 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*

*Lowest prediction errors.

levels relative to band edges, while the size of the lattice and heat of fusion may determine how likely it is for the defect atom to exist at site in consideration

To explore linear relationships further, we chose multiple linear regression (MLR) as the method to train the first predictive models. Given a vector of properly standardized features $X^T = (X_1, X_2, \ldots, X_p)$, and calculated output vector Y, the matrix of coefficients β corresponding with each feature and the 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, least absolute shrinkage and selection operator (LASSO) regression, ridge regression, and elastic net regression,[78] all of which yield a biased predictor, but with a lower variance, leading to less overfitting compared with the standard least square. Ridge regression shrinks the coefficients β by imposing an L_2 penalty, whereas LASSO uses an L_1 penalty.[79] The elastic net is another regularized linear regression technique that combines both L_1 penalty and L_2 penalty. Typically, β shrinkage inside LASSO regression progress more severely compared with the other two approaches, and some of the coefficients are brought down to 0.

In Figures S4–S7, 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 1 and 2 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 nonlinear 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 nonlinear 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 toward II–VI semiconductors and interstitial site impurities, as seen from Figure 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 nonlinear 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 nonlinear 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 sub-sample of the dataset, chosen through bootstrapping. During the construction of any tree, the best split for each node is found based on some number of input features. Averaging 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 to control overfitting.[80]

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 performed[81] 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. Figure S9 shows a comparison between grid search and Bayesian search based hyperparameter optimization for RFR; it is seen that both methods produce similar test errors, but the latter mitigates overfitting (difference between training and test errors) far better, thus motivating its use. Parity plots for the optimized models for all eight properties are shown in Figure 5A. Looking at the error values listed in Tables 1 and 2, 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 and 0.38 eV for RFR, and the formation energy RMSEs drop from 1.2 eV or

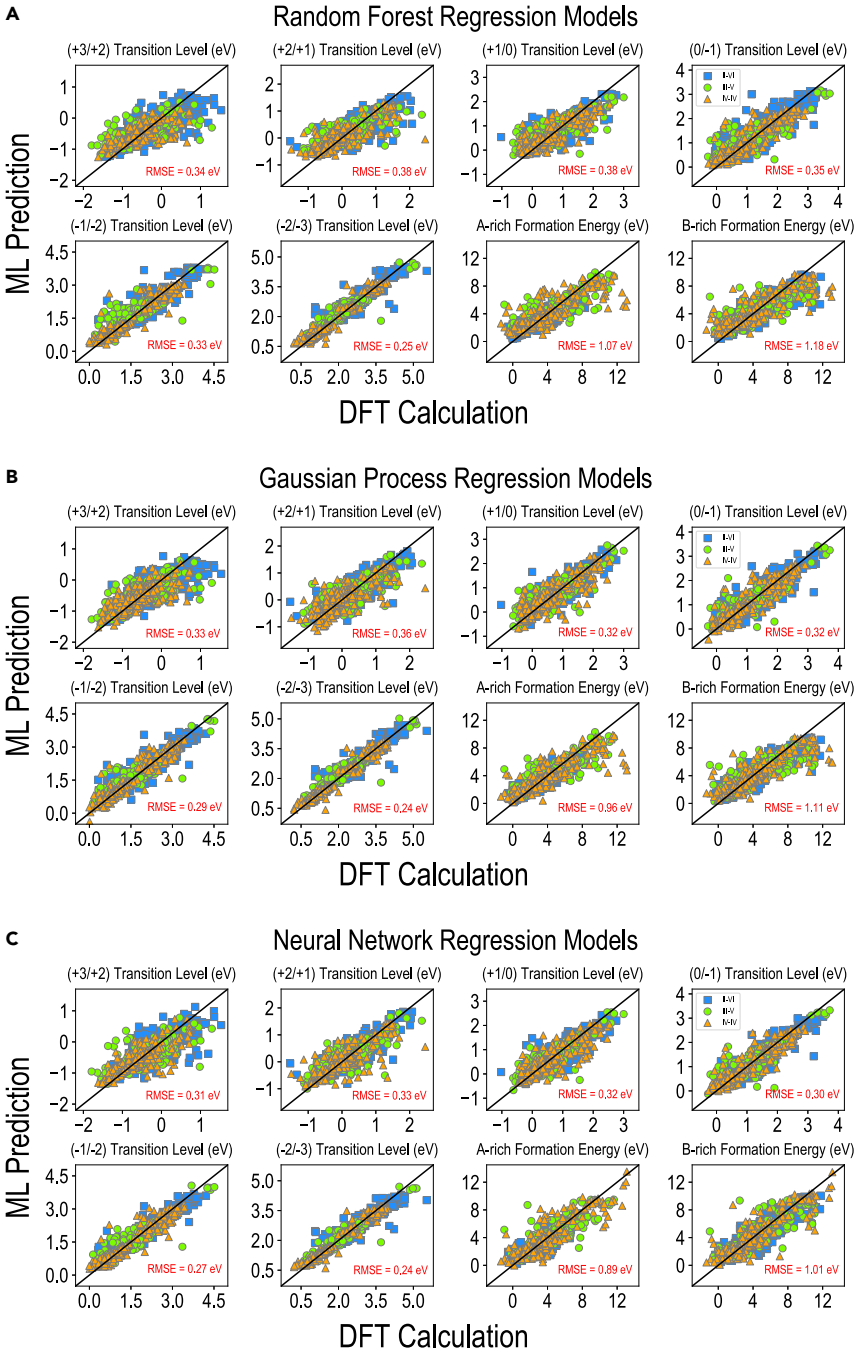

Figure 5. Parity plots for best regression models

(A–C) (A) Random Forest, (B) Gaussian process, and (C) NN regression, plotted for different semiconductor types.

higher to between 1.07 and 1.16 eV. We further plotted the RFR uncertainties in prediction against the absolute prediction error values in Figure S8. 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 III–V or IV–IV, owing to the larger portion of II–VI semiconductor points in the training dataset. 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 dataset 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.

Finally, we examined the feature importance values that are reported as part of random forest models; these values were collected for each property and averaged over five-fold cross validation. The importance values for the ten best (SISSO-based) features are plotted for all eight properties in Figure S10. It is seen that the most important features for predicting $\varepsilon(+3/+2)$, $\varepsilon(+2/+1)$ and $\varepsilon(+1/0)$ are the differences between valence, preferred oxidation states, and electronegativities of the defect atom M and A/B atoms. The $\varepsilon(0/-1)$ is determined by the difference between thermal expansion coefficient of M and thermal conductivity of B, $\varepsilon(-1/-2)$ by the atomic radius of B and electrical conductivity of A, and $\varepsilon(-2/-3)$ by the covalent radius of B and Mendeleev number of A. While many of the important features for transition levels are similar to those obtained from Figure S3, the emergence of some new features shows the importance of uncovering more complex nonlinear relationships. Finally, the most important features for ΔH (A rich) and ΔH (B rich) are differences between group numbers and heat of vaporization for the former and differences in thermal expansion coefficients for the latter.

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 that uses the kernel trick to

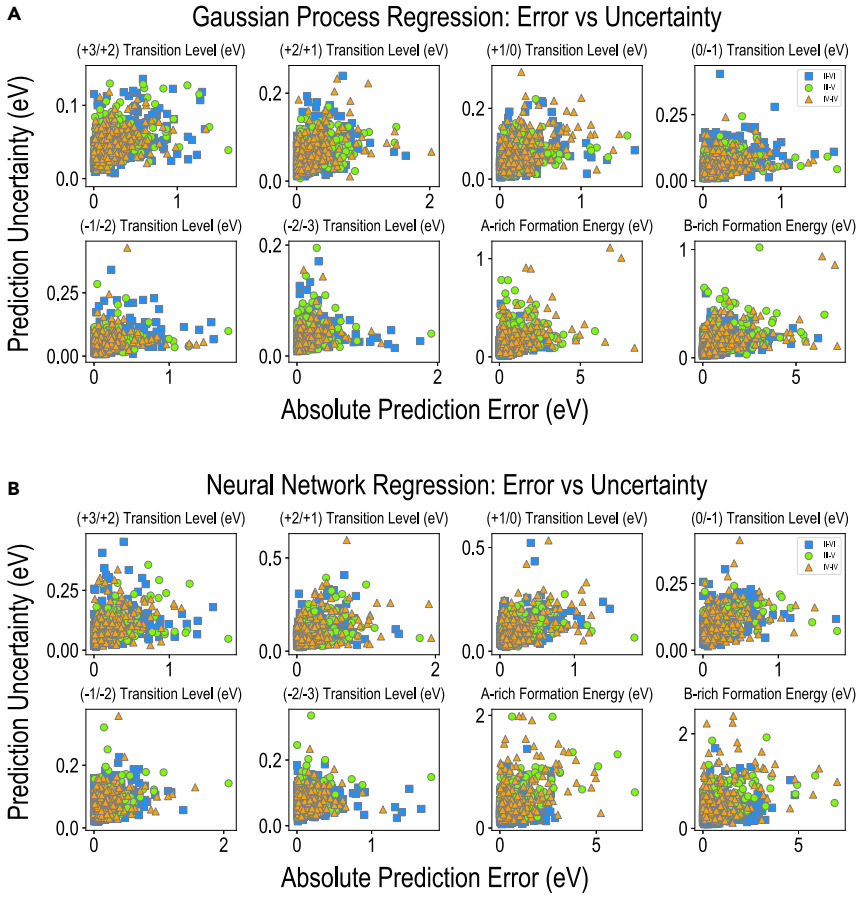

Figure 6. Gaussian process regression: Error versus uncertainty

(A and B) Prediction uncertainty as a function of absolute prediction error for (A) Gaussian process and (B) NN regression, plotted for different semiconductor types.

solve a nonlinear problem in a linear fashion. The original low-dimensional features are used as input and mapped to a high-dimensional kernel space in which they can be linearly interpreted. In this work, we use different possible choices for the kernel function, namely polynomial, radial basis function, 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 Figure S11A and listed in Tables 1 and 2. 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 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 six transition levels range between 0.25 and 0.35 eV. As shown in Figure S11B, 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 nonlinear regression technique that uses the kernel trick is Gaussian process regression (GPR).[82,83] 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' theorem,[84] 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 Figure 5B and from Tables 1 and 2 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 S6 and S7 that there is a greater difference between the training and the 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 features.[85] The uncertainty versus absolute error plots in Figure 6A 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 feedforward NN to machine learn a multi-output regression.[86] 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 used 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 used 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 each dense layer, beating out sigmoid, softmax, softplus, tanh, and selu functions,[87] while the Adam optimizer was selected over SGD, RMSEprop, Adadelta, and Adagrad.[88] NN model training involved 10-repeated 5-fold cross-validation, where the mean and SD 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 Figure 5C, while Figure 6B shows the uncertainty versus absolute error plots.

It can be seen from the parity plots and Tables 1 and 2 that NN predictions for both transition levels and formation energies are similar to KRR and GPR. Transition level RMSE values are 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 S6 and S7 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 Figure 6B, 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 SD in prediction; this is expected to affect NN predictions over the entire chemical space. We further note that standard deviations over 10-folds may not be sufficient to converge the uncertainties, but we use an ensemble of 10 predictions here to save on training time and keep estimates consistent across different ML models. Methods such as Monte Carlo dropout[89] can help to attain better uncertainty estimates as well, and will be applied in future work.

High-throughput screening of dominating impurities

The detailed ML 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 Figure 7A, we present the test set prediction RMSE values of eight 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 nonlinear 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. It is important to note here that a significant amount of time is saved by replacing full DFT calculations with almost instantaneous ML predictions. On average, any 1 point defect in a 64-atom supercell simulated in the neutral state requires approximately 500 core hours, while 6 charged state calculations require a further approximately 2000 core hours (running on 8 Intel Broadwell XEON E5-2695 nodes with 36 cores each). For the DFT datasets of approximately 1500 neutral state formation energies and approximately 1000 charge transition levels of 6 types, this translates to approximately 2.75 million core hours. For the entire dataset of 12,474 impurities, approximately 32 million core hours would be required for complete DFT optimization and prediction of all defect properties. In contrast, every ML model takes a matter of minutes to train and make predictions over the entire chemical space. Thus, based on computations using 1/10th of the total computing time required, we can make reasonable predictions for all the data points. Predictions for the entire set of 12,474 impurities using different ML models are included as a spreadsheet as part of the Supplementary materials.

The ML-predicted impurity formation energies across the dataset were 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 versus experiments and ML versus 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:

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN)$$

$$\text{Precision} = TP/(TP + FP)$$

$$\text{Recall} = TP/(TP + FN)$$

Figures 7B and 7C 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 Figure 7B, 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 uncertainty values seen in Figure 6B 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 Figure 7C,

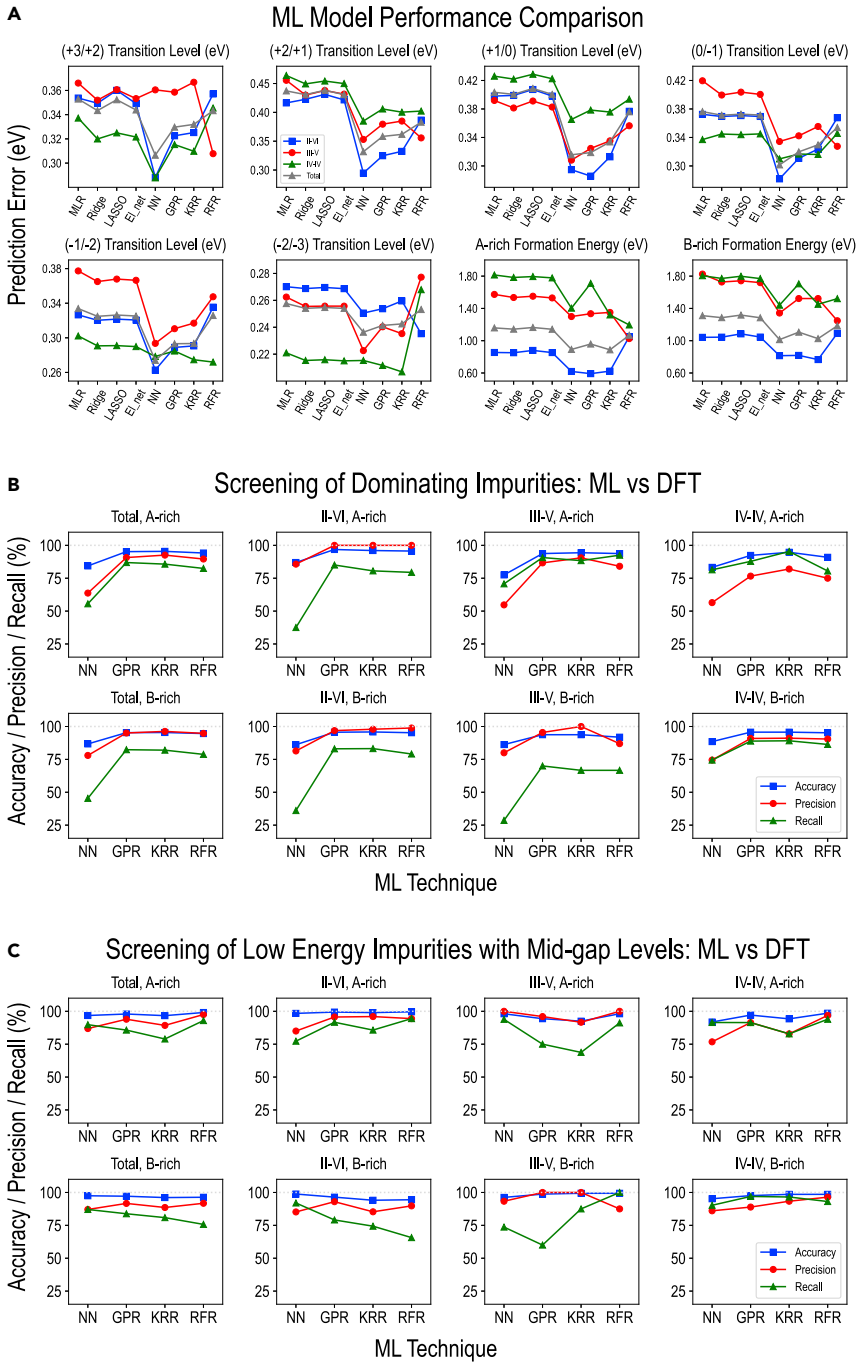

Figure 7. ML model performance comparison

(A–C) 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 at A-rich and B-rich chemical potential conditions.

and all four techniques (NN, GPR, KRR, and RFR) 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. To further elucidate the effect of this dataset imbalance, we retrained some GPR and RFR models on a reduced dataset with (a) equal number of II–VI, III–V, and IV–IV points, and (b) equal number of interstitial and substitutional points, essentially by removing a large number of II–VI or interstitial points (mostly belonging to impurities in CdTe). Figures S15 and S16 show, respectively, the GPR and RFR RMSE values for all properties using the entire dataset and using reduced balanced datasets. It can be seen that the overall predictions are slightly worse with the balanced datasets, simply because fewer data points are being used for training, but the gap between II–VI, III–V, and IV–IV errors reduces, as does the gap between substitutional and interstitial errors. However, for formation energy, the II–VI points and interstitial points are still predicted better than other data types. We conclude that, although errors for different data types can be brought closer to each other using more balanced datasets, we prefer to use models trained on the entire dataset since they lead to similar or better errors for different data types.

In Table 3, we list 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 Figure 8. 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 Figure 8 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 with DFT, which implies a reliable qualitative screening, even when the actual predicted formation energies or transition levels are off. As a final test of the generalizability of our ML framework, we selected 25 new impurities deemed to be dominating from GPR and

Table 3. 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

KRR predictions, and performed additional computations on them. Figure S14 shows the parity plots between the DFT-computed formation energies and transition levels and the GPR/KRR-predicted values. It can be seen that RMSE values are generally between 0.8 and 1.1 eV for formation energies and between approximately 0.2 and 0.4 eV for the transition levels, indicating that prediction accuracy is at a very similar level to test set predictions.

DISCUSSION

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 greater than 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 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.[90,91]

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.

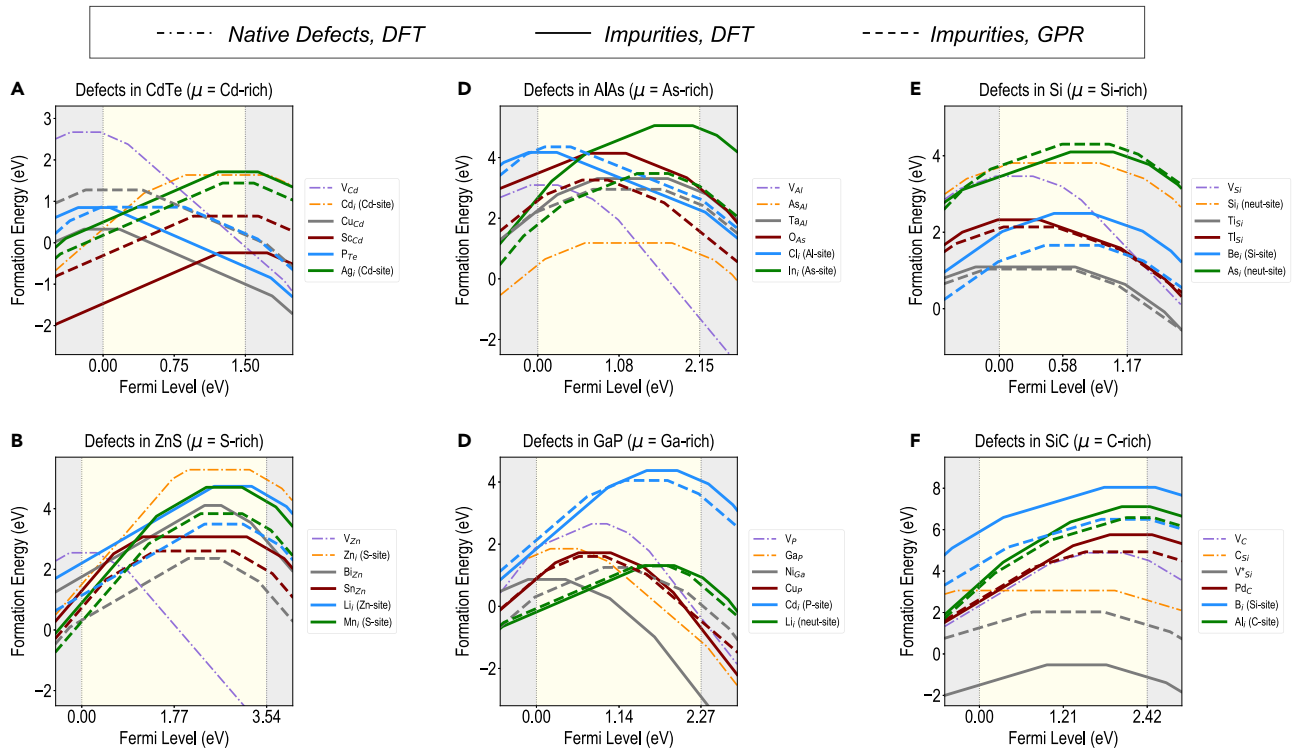

Figure 8. Defect formation energies from DFT and ML

(A–F) 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. The dominant donor and acceptor type native defects are also pictured.

In ref[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 versus experiment gap. Further, while all compounds were currently studied in the ZB structure, the DFT data and ML models will be extended to include defects in the wurtzite and rock salt 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-chalcogenides.[12] 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 refs.[3,12] Finally, tools can be created for the 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.

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 ZB 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 upward 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 NN. For the eight 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.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Requests for data and additional information should be directed to the lead contact, Arun Mannodi-Kanakkithodi (amannodi@purdue.edu).

Materials availability

This study did not generate physical materials.

Data and code availability

DFT data and ML models are made available as part of an open-source tool to facilitate the artificial intelligence-driven prediction and screening of point defects and impurities in semiconductors: https://github.com/lmjacoby/ai_semiconductors.

Calculating defect properties from DFT

All native defects and impurities were simulated in 64 atom 2 × 2 × 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[92,93] using the PBE[94] exchange-correlation functional and projector-augmented wave atom potentials.[95] The kinetic energy cut-off for the plane-wave 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 × 3 × 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, charge q, and Fermi level E_F, and any impurity charge transition level, $\varepsilon(q_1/q_2)$:

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

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

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 VBM 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.[96,97] 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 versus the 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 Equation 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. Equation 2 defines a charge transition level $\varepsilon(q_1/q_2)$, that is, 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.

ML details

The ML approaches used in this work include dimensionality reduction/outlier identification using SISSO, PCA, and other techniques, and training predictive models using linear regression and three types of nonlinear regression: random forests, Gaussian processes, and NN. Necessary introduction to each technique and relevant information about how hyperparameters are optimized and errors are converged are provided in different subsections within the manuscript. All ML training and prediction was done using appropriate functions in Scikit-learn.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100450.

ACKNOWLEDGMENTS

This work was performed in part at the Center for Nanoscale Materials, a US Department of Energy Office of Science User Facility, and supported by the US 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 US 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. and X.X. acknowledge 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.

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.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 11, 2021
Revised: December 6, 2021
Accepted: January 20, 2022
Published: February 14, 2022

REFERENCES

1. Guan, Z., Chen, F., Liu, Z., Lv, P., Chen, M., Guo, M., et al. (2019). Compositional engineering of multinary Cu–In–Zn-based semiconductor nanocrystals for efficient and solution-processed red-emitting quantum-dot light-emitting diodes. Org. Electron. *74*, 46–51.
2. Klug, M.T., Osherov, A., Haghighirad, A.A., Stranks, S.D., Brown, P.R., Bai, S., Wang, J.T.-W., Dang, X., Bulovic, V., Snaith, H.J., and Belcher, A.M. (2017). Tailoring metal halide perovskites through metal substitution: influence on photovoltaic and material properties. Energy Environ. Sci. *10*, 236–246.
3. Mannodi-Kanakkithodi, A., Park, J.-S., Jeon, N., Cao, D.H., Gosztola, D.J., Martinson, A.B.F., and Chan, M.K.Y. (2019). Comprehensive computational study of partial lead substitution in methylammonium lead bromide. Chem. Mater. *31*, 3599–3612.
4. Ning, C.-Z., Dou, L., and Yang, P. (2017). Bandgap engineering in semiconductor alloy nanomaterials with widely tunable compositions. Nat. Rev. Mater. *2*, 1–14.
5. Oba, F., and Kumagai, Y. (2018). Design and exploration of semiconductors from first principles: a review of recent advances. Appl. Phys. Express *11*, 060101.
6. Sampson, M.D., Park, J.S., Schaller, R.D., Chan, M.K.Y., and Martinson, A.B.F. (2017). Transition metal-substituted lead halide perovskite absorbers. J. Mater. Chem. A. *5*, 3578–3588.
7. Krasikov, D., and Sankin, I. (2017). Defect interactions and the role of complexes in the CdTe solar cell absorber. J. Mater. Chem. A. *5*, 3503–3513.
8. Park, J.S., Kim, S., Xie, Z., and Walsh, A. (2018). Point defect engineering in thin-film solar cells. Nat. Rev. Mater. 2058–8437.
9. Queisser, H.J., and Haller, E.E. (1998). Defects in semiconductors: some fatal, some vital. Science *281*, 945–950.

10. Yin, W.-J., Shi, T., and Yan, Y. (2014). Unusual defect physics in $CH_3NH_3PbI_3$ perovskite solar cell absorber. Appl. Phys. Lett. *104*, 063903.
11. Mannodi-Kanakkithodi, A., Park, J.-S., Martinson, A.B.F., and Chan, M.K.Y. (2020). Defect energetics in pseudo-cubic mixed halide lead perovskites from first-principles. J. Phys. Chem. C *124*, 16729–16738.
12. Mannodi-Kanakkithodi, A., Toriyama, M.Y., Sen, F.G., Davis, M.J., Klie, R.F., and Chan, M.K. (2020). Machine-learned impurity level prediction for semiconductors: the example of Cd-based chalcogenides. NPJ Comput. Mater. *6*, 1–14.
13. Liu, Z., Na, G., Tian, F., Yu, L., Li, J., and Zhang, L. (2020). Computational functionality-driven design of semiconductors for optoelectronic applications. InfoMat *2*, 879–904.
14. Nayak, P.K., Mahesh, S., Snaith, H.J., and Cahen, D. (2019). Photovoltaic solar cell technologies: analysing the state of the art. Nat. Rev. Mater. *4*, 269–285.
15. Sivathanu, V., Thangavel, R., and Lenka, T.R. (2021). Modeling and performance optimization of two-terminal Cu_2ZnSnS_4–Silicon tandem solar cells. Int. J. Energy Res. *46*, 104–123.
16. Torres-Jaramillo, S., Bernal-Correa, R., and Morales-Acevedo, A. (2021). Improved design of InGaP / GaAs / Si tandem solar cells. EPJ Photovolt *12*, 1.
17. Wang, J., Chen, H., Wei, S.-H., and Yin, W.-J. (2019). Materials design of solar cell absorbers beyond perovskites and conventional semiconductors via combining tetrahedral and octahedral coordination. Adv. Mater. *31*, 1806593.
18. Adachi, S. (2005). Energy-Band Structure: Energy-Band Gaps, Chapter 6 (John Wiley and Sons, Ltd.), pp. 103–145.
19. Bouarissa, N., Algarni, H., Ajmal Khan, M., Al-Hagan, O., and Alhuwaymel, T. (2020). Collective effects and optical characteristics of $CdSe_xTe_{1-x}$. Optik *203*, 163952.
20. Dumre, B., Szymanski, N., Adhikari, V., Khatri, I., Gall, D., and Khare, S. (2019). Improved optoelectronic properties in $CdSe_xTe_{1-x}$ through controlled composition and short-range order. Solar Energy *194*, 742–750.
21. Ban, D., Luo, H., Liu, H.C., Wasilewski, Z.R., SpringThorpe, A.J., Glew, R., and Buchanan, M. (2004). Optimized gaasalgaas light-emitting diodes and high efficiency wafer-fused optical up-conversion devices. J. Appl. Phys. *96*, 5243–5248.
22. Wang, G., Yi, X., Zhan, T., and Huang, Y. (2019). The AlGaInP / AlGaAs Material System and Red / Yellow LED (Springer International Publishing), pp. 171–202.
23. Heo, S., Seo, G., Lee, Y., Lee, D., Seol, M., Lee, J., et al. (2017). Deep level trapped defect analysis in $CH_3NH_3PbI_3$ perovskite solar cells by deep level transient spectroscopy. Energy Environ. Sci. *10*, 1128–1133.
24. Rosenberg, J.W., Legodi, M.J., Rakita, Y., Cahen, D., and Diale, M. (2017). Laplace current deep level transient spectroscopy measurements of defect states in methylammonium lead bromide single crystals. J. Appl. Phys. *122*, 145701.
25. Schultz, P.A. (2006). Theory of defect levels and the "band gap problem" in Silicon. Phys. Rev. Lett. *96*, 246401.
26. Mannodi-Kanakkithodi, A., and Chan, M.K. (2021). Computational data-driven materials discovery. Trends Chem. *3*, 79–82, Special issue: machine learning for Molecules and materials.
27. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., and Kim, C. (2017). Machine learning in materials informatics: recent applications and prospects. NPJ Comput. Mater. *3*, 54.
28. Schmidt, J., Marques, M.R.G., Botti, S., and Marques, M.A.L. (2019). Recent advances and applications of machine learning in solid-state materials science. NPJ Comput. Mater. *5*, 83.
29. Agrawal, A., and Choudhary, A. (2018). An online tool for predicting fatigue strength of steel alloys based on ensemble data mining. Int. J. Fatigue *113*, 389–400.
30. Kim, C., Chandrasekaran, A., Huan, T.D., Das, D., and Ramprasad, R. (2018). Polymer genome: a data-powered polymer informatics platform for property predictions. J. Phys. Chem. C *122*, 17575–17585.
31. Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N.E., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., et al. (2018). Matminer: an open source toolkit for materials data mining. Comput. Mater. Sci. *152*, 60–69.
32. Doan, H.A., Agarwal, G., Qian, H., Counihan, M.J., Rodríguez-López, J., Moore, J.S., and Assary, R.S. (2020). Quantum chemistry-informed active learning to accelerate the design and discovery of sustainable energy storage materials. Chem. Mater. *32*, 6338–6346.
33. Feng, H.-J., Wu, K., and Deng, Z.-Y. (2020). Predicting inorganic photovoltaic materials with efficiencies >26% via structure-relevant machine learning and density functional calculations. Cell Rep. Phys. Sci. *1*, 100179.
34. Iwasaki, Y., Takeuchi, I., Stanev, V., Kusne, A.G., Ishida, M., Kirihara, A., Ihara, K., Sawada, R., Terashima, K., Someya, H., et al. (2019). Machine-learning guided discovery of a new thermoelectric material. Sci. Rep. *9*, 2751.
35. Khmaissia, F., Frigui, H., Sunkara, M., Jasinski, J., Garcia, A.M., Pace, T., and Menon, M. (2018). Accelerating band gap prediction for solar materials using feature selection and regression techniques. Comput. Mater. Sci. *147*, 304–315.
36. Mannodi-Kanakkithodi, A., Chandrasekaran, A., Kim, C., Huan, T.D., Pilania, G., Botu, V., and Ramprasad, R. (2018). Scoping the polymer genome: a roadmap for rational polymer dielectrics design and beyond. Mater. Today *21*, 785–796.
37. Meredig, B., Agrawal, A., Kirklin, S., Saal, J.E., Doak, J.W., Thompson, A., Zhang, K., Choudhary, A., and Wolverton, C. (2014). Combinatorial screening for new materials in unconstrained composition space with machine learning. Phys. Rev. B *89*, 094104.
38. Vasudevan, R.K., Choudhary, K., Mehta, A., Smith, R., Kusne, G., Tavazza, F., Vlcek, L., Ziatdinov, M., Kalinin, S.V., and Hattrick-Simpers, J. (2019). Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics. MRS Commun. *9*, 821–838.
39. Deml, A.M., Holder, A.M., O'Hayre, R.P., Musgrave, C.B., and Stevanović, V. (2015). Intrinsic material properties dictating oxygen vacancy formation energetics in metal oxides. J. Phys. Chem. Lett. *6*, 1948–1953.
40. Deml, A.M., Stevanović, V., Muhich, C.L., Musgrave, C.B., and O'Hayre, R. (2014). Oxide enthalpy of formation and band gap energy as accurate descriptors of oxygen vacancy formation energetics. Energy Environ. Sci. *7*, 1996–2004.
41. Sharma, V., Kumar, P., Dev, P., and Pilania, G. (2020). Machine learning substitutional defect formation energies in ABO_3 perovskites. J. Appl. Phys. *128*, 034902.
42. Wan, Z., Wang, Q.-D., Liu, D., and Liang, J. (2021). Data-driven machine learning model for the prediction of oxygen vacancy formation energy of metal oxide materials. Phys. Chem. Chem. Phys. *23*, 15675–15684.
43. Frey, N.C., Akinwande, D., Jariwala, D., and Shenoy, V.B. (2020). Machine learning-enabled design of point defects in 2D materials for quantum and neuromorphic information processing. ACS Nano *14*, 13406–13417.
44. Varley, J.B., Samanta, A., and Lordi, V. (2017). Descriptor-based approach for the prediction of cation vacancy formation energies and transition levels. J. Phys. Chem. Lett. *8*, 5059–5063.
45. Miller, S.A., Dylla, M., Anand, S., Gordiz, K., Snyder, G.J., and Toberer, E.S. (2018). Empirical modeling of dopability in diamond-like semiconductors. NPJ Comput. Mater. *4*, 71.
46. Ramprasad, R., Zhu, H., Rinke, P., and Scheffler, M. (2012). New perspective on formation energies and energy levels of point defects in nonmetals. Phys. Rev. Lett. *108*, 066404.
47. Aryasetiawan, F., and Gunnarsson, O. (1998). The GW method. Rep. Prog. Phys. *61*, 237–312.
48. Chan, M., and Ceder, G. (2010). Efficient band gap prediction for solids. Phys. Rev. Lett. *105*, 196403.

49. Heyd, J., Peralta, J.E., Scuseria, G.E., and Martin, R.L. (2005). Energy band gaps and lattice parameters evaluated with the Heyd-Scuseria-Ernzerhof screened hybrid functional. J. Chem. Phys. *123*, 174101.
50. Schultz, P.A. (2011). First principles predictions of intrinsic defects in Aluminum Arsenide, AlAs. MRS Proc. *1370*, mrss11–1370–yy03–04.
51. Ablekim, T., Swain, S.K., Yin, W.-J., Zaunbrecher, K., Burst, J., Barnes, T.M., et al. (2017). Self-compensation in arsenic doping of CdTe. Sci. Rep. *7*, 4563.
52. Ayoub, M., Hage-Ali, M., Koebel, J.M., Zumbiehl, A., Klotz, F., Rit, C., et al. (2003). Annealing effects on defect levels of CdTe:Cl materials and the uniformity of the electrical properties. IEEE Trans. Nucl. Sci. *50*, 229–237.
53. Chen, J.W., and Milnes, A.G. (1980). Energy levels in Silicon. Annu. Rev. Mater. Sci. *10*, 157–228.
54. Dow, J.D., Ren, S.Y., and Shen, J. (1988). Deep Impurity Levels in Semiconductors, Semiconductor Alloys, and Superlattices (Springer US), pp. 175–187.
55. Grimmeiss, H.G. (1977). Deep level impurities in semiconductors. Annu. Rev. Mater. Sci. *7*, 341–376.
56. Itoh, H., Kawasuso, A., Ohshima, T., Yoshikawa, M., Nashiyama, I., Tanigawa, S., et al. (1997). Intrinsic defects in cubic Silicon Carbide. Phys. Status Solidi (A) *162*, 173–198.
57. Itoh, H., Yoshikawa, M., Nashiyama, I., Okumura, H., Misawa, S., and Yoshida, S. (1995). Photoluminescence of radiation induced defects in 3C–SiC epitaxially grown on Si. J. Appl. Phys. *77*, 837–842.
58. Jantsch, W., and Hendorfer, G. (1990). Characterization of deep levels in CdTe by photo-epr and related techniques. J. Cryst. Growth *101*, 404–413.
59. Komin, V., Viswanathan, V., Tetali, B., Morel, D.L., and Ferekides, C.S. (2000). Investigation of deep levels in CdTe / CdS solar cells. In Conference Record of the Twenty-Eighth IEEE Photovoltaic Specialists Conference - 2000 (Cat. No.00CH37036) (IEEE), pp. 676–679.
60. Kraft, C., Brömel, A., Schönherr, S., Hädrich, M., Reislöhner, U., Schley, P., et al. (2011). Phosphorus implanted Cadmium Telluride solar cells. Thin Solid Films *519*, 7153–7155, Proceedings of the EMRS 2010 spring meeting symposium M: thin film chalcogenide photovoltaic materials.
61. Lebedev, A.A. (1999). Deep level centers in Silicon Carbide: a review. Semiconductors *33*, 107–130.
62. Lindström, A., Mirbt, S., Sanyal, B., and Klintenberg, M. (2015). High resistivity in undoped CdTe: carrier compensation of Te antisites and Cd vacancies. J. Phys. D Appl. Phys. *49*, 035101.
63. Nagesh, V., Farmer, J.W., Davis, R.F., and Kong, H.S. (1990). Defects in cubic SiC on Si. Radiat. Effects Defects Solids *112*, 77–84.
64. Schöler, M., Lederer, M.W., Schuh, P., and Wellmann, P.J. (2020). Intentional incorporation and tailoring of point defects during sublimation growth of cubic Silicon Carbide by variation of process parameters. Phys. Status Solidi (b) *257*, 1900286.
65. Swaminathan, V. (1982). Defects in GaAs. Bull. Mater. Sci. *4*, 403–442.
66. Zhou, P., Spencer, M.G., Harris, G.L., and Fekade, K. (1987). Observation of deep levels in cubic Silicon Carbide. Appl. Phys. Lett. *50*, 1384–1385.
67. Kononova, O., He, T., Huo, H., Trewartha, A., Olivetti, E.A., and Ceder, G. (2021). Opportunities and challenges of text mining in materials research. iScience *24*, 102155.
68. Olivetti, E.A., Cole, J.M., Kim, E., Kononova, O., Ceder, G., Han, T.Y.-J., and Hiszpanski, A.M. (2020). Data-driven materials research enabled by natural language processing and information extraction. Appl. Phys. Rev. *7*, 041317.
69. Sun, R., Chan, M.K.Y., Kang, S., and Ceder, G. (2011). Intrinsic stoichiometry and Oxygen-induced *p*-type conductivity of Pyrite FeS_2. Phys. Rev. B *84*, 035212.
70. FElton, D.C., Boukouvalas, Z., Butrico, M.S., Fuge, M.D., and Chung, P.W. (2018). Applying machine learning techniques to predict the properties of energetic materials. Sci. Rep. *8*, 9059.
71. Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M., and Ghiringhelli, L.M. (2018). Sisso: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. Phys. Rev. Mater. *2*, 083802.
72. Fan, J., and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. Ser. B (Stat. Methodol.) *70*, 849–911.
73. Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. J. Mach. Learn. Res. *10*, 2013–2038.
74. Kramer, O. (2013). K-nearest Neighbors (Springer), pp. 13–23.
75. George, A., and Vidyapeetham, A. (2012). Anomaly detection based on machine learning: dimensionality reduction using PCA and classification using SVM. Int. J. Computer Appl. *47*, 5–8.
76. Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., and Chang, L. (2003). 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) (IEEE), pp. 172–179.
77. Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., and Mueller, A. (2015). Scikit-learn: machine learning without learning the machinery. Getmobile Mobile Comp. Commun. *19*, 29–33.
78. Friedman, J., Hastie, T., and Tibshirani, R. (2001). The Elements of Statistical Learning, *1* (Springer series in statistics).
79. Ng, A.Y. (2004). Feature selection, l1 vs. l2 regularization, and rotational invariance. In Proceedings of the Twenty-First International Conference on Machine Learning, ICML '04 (Association for Computing Machinery), p. 78.
80. Breiman, L. (2001). Random forests. Machine Learn. *45*, 5–32.
81. Bergstra, J., Yamins, D., and Cox, D. (2013). 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, *28* (JMLR).
82. Seeger, M. (2004). Gaussian processes for machine learning. Int. J. Neural Syst. *11*, 69–106.
83. Williams, C.K., and Rasmussen, C.E. (2006). Gaussian Processes for Machine Learning, *2* (MIT press Cambridge).
84. Puga, J.L., Krzywinski, M., and Altman, N. (2015). Bayes' theorem *12*, 277–278.
85. Mohammed, R.O., and Cawley, G.C. (2017). Over-fitting in model selection with Gaussian process regression. In International Conference on Machine Learning and Data Mining in Pattern Recognition (Springer), pp. 192–205.
86. Bisong, E. (2019). Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners (Apress).
87. Ramachandran, P., Zoph, B., and Le, Q.V. (2018). Searching for Activation Functions (International Conference on Learning Representations).
88. Choi, D., Shallue, C.J., Nado, Z., Lee, J., Maddison, C.J., and Dahl, G.E. (2020). On Empirical Comparisons of Optimizers for Deep Learning (International Conference on Learning Representations).
89. Camarasa, R., Bos, D., Hendrikse, J., Nederkoorn, P., Kooi, E., van der Lugt, A., and de Bruijne, M. (2020). Quantitative comparison of monte-carlo dropout uncertainty measures for multi-class segmentation. In Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis (Springer International Publishing), pp. 32–41.
90. Fazzio, A., Caldas, M.J., and Zunger, A. (1984). Many-electron multiplet effects in the spectra of 3*d* impurities in heteropolar semiconductors. Phys. Rev. B *30*, 3430–3455.
91. Sugano, S., Tanabe, Y., and Kamimura, H. (1970). Multiplets of Transition-Metal Ions in Crystals (Academic Press).

92. Kresse, G., and Furthmüller, J. (1996). Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. Phys. Rev. B *54*, 11169–11186.

93. Kresse, G., and Hafner, J. (1994). Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in Germanium. Phys. Rev. B *49*, 14251–14269.

94. Perdew, J.P., Burke, K., and Ernzerhof, M. (1996). Generalized gradient approximation made simple. Phys. Rev. Lett. *77*, 3865–3868.

95. Blöchl, P.E. (1994). Projector augmented-wave method. Phys. Rev. B *50*, 17953–17979.

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

97. Freysoldt, C., Neugebauer, J., and Van de Walle, C.G. (2009). Fully ab initio finite-size corrections for charged-defect supercell calculations. Phys. Rev. Lett. *102*, 016402.

Patterns, Volume *3*

Supplemental information

Universal machine learning framework for defect predictions in zinc blende semiconductors

Arun Mannodi-Kanakkithodi, Xiaofeng Xiang, Laura Jacoby, Robert Biegaj, Scott T. Dunham, Daniel R. Gamelin, and Maria K.Y. Chan

Supplemental Information

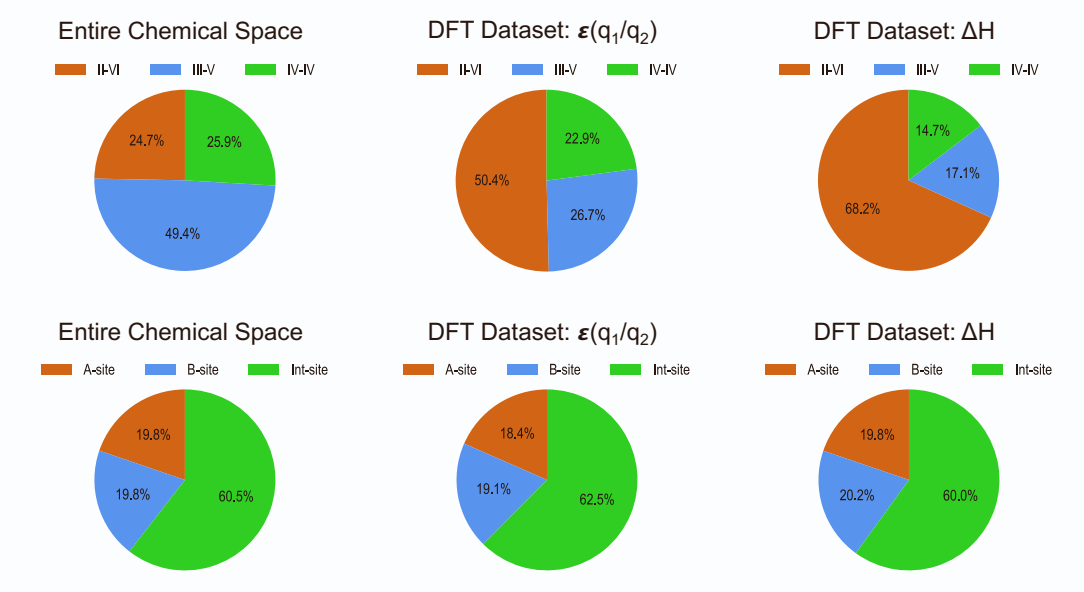

FIG. S1: Pie charts visualizing the distribution of different semiconductor and defect site types in the dataset.

TABLE SI: Reported experimental values and calculated results from PBE, HSE and HSE+SOC for the lattice constants and band gaps of all semiconductor compounds used in this work.

Type	**Compound**	**Expt. Lattice Constant (Å)**	**PBE Lattice Constant (Å)**	**HSE Lattice Constant (Å)**	**Expt. Band Gap (eV)**	**PBE Band Gap (eV)**	**HSE Band Gap (eV)**	**HSE+SOC Band Gap (eV)**
II-VI	CdO	5.10	5.15	5.10	2.30	0.47	0.99	0.97
II-VI	CdS	5.83	5.94	5.90	2.42	1.03	2.10	2.08
II-VI	CdSe	6.08	6.21	6.15	1.70	0.48	1.48	1.34
II-VI	CdTe	6.48	6.62	6.57	1.50	0.57	1.43	1.14
II-VI	ZnO	4.60	4.63	4.58	3.27	0.62	2.35	2.33
II-VI	ZnS	5.42	5.45	5.43	3.54	2.01	3.28	3.25
II-VI	ZnSe	5.67	5.74	5.70	2.70	1.14	2.33	2.19
II-VI	ZnTe	6.10	6.18	6.14	2.26	1.06	2.05	1.75
III-V	BN	3.62	3.63	3.60	6.10	4.44	5.81	-
III-V	BP	4.54	4.55	4.52	2.10	1.24	1.96	-
III-V	BAs	4.78	4.82	4.76	1.82	1.20	1.84	-
III-V	BSb	5.10	5.28	5.20	0.55	0.75	1.24	-
III-V	AlN	4.38	4.40	4.36	5.30	3.30	4.58	-
III-V	AlP	5.45	5.51	5.47	2.45	1.62	2.30	-
III-V	AlAs	5.66	5.73	5.69	2.15	1.49	2.13	-
III-V	AlSb	6.14	6.23	6.14	1.62	1.22	1.84	-
III-V	GaN	4.52	4.59	4.52	3.20	1.45	2.98	-
III-V	GaP	5.45	5.53	5.45	2.27	1.51	2.30	-
III-V	GaAs	5.65	5.77	5.67	1.42	0.13	1.34	-
III-V	GaSb	6.10	6.22	6.10	0.75	0.00	0.91	-
III-V	InN	4.98	5.10	5.03	0.70	0.00	0.46	-
III-V	InP	5.87	6.00	5.93	1.34	0.37	1.33	-
III-V	InAs	6.06	6.22	6.11	0.36	0.00	0.38	-
III-V	InSb	6.48	6.65	6.52	0.17	0.00	0.42	-
IV-IV	C	3.57	3.57	3.54	5.50	4.12	-	-
IV-IV	Si	5.43	5.47	5.45	1.17	0.61	-	-
IV-IV	Ge	5.65	5.79	5.68	0.66	0.00	0.70	-
IV-IV	Sn	6.49	6.65	6.54	0.12	0.00	0.00	-
IV-IV	SiC	4.36	4.38	4.35	2.42	1.36	2.25	-
IV-IV	GeC	4.60	4.63	4.57	1.42	1.63	2.39	-
IV-IV	SnC	5.02	5.10	5.03	1.22	0.59	1.65	-
IV-IV	SiGe	5.54	5.62	5.55	0.92	0.61	1.12	-
IV-IV	SiSn	-	6.09	6.01	-	0.34	1.09	-
IV-IV	GeSn	-	6.23	6.12	-	0.00	0.13	-
-								

TABLE SII: Experimentally measured defect levels (collected from references[51–66]) and their corresponding DFT-PBE computed values, as plotted in Fig. 2. All levels are relative to the semiconductor valence band maximum.

Compound	**Semiconductor Type**	**Defect**	**Defect Transition Label**	**Measured Value**	**DFT-PBE Value**
CdTe	II-VI	V_{Cd}	(-1/-2)	0.43	0.32
CdTe	II-VI	As_{Te}	(0/-1)	0.09	0.07
CdTe	II-VI	Cu_{Cd}	(0/-1)	0.35	0.15
CdTe	II-VI	Cu_i	(+1/0)	1.05	1.04
CdTe	II-VI	Te_{Cd}	(+2/+1)	0.32	0.35
CdTe	II-VI	Fe_{Cd}	(+1/0)	0.20	0.23
CdTe	II-VI	Cd_i	(+2/+1)	0.64	0.78
CdTe	II-VI	V^*_{Cd}	(+1/0)	0.65	0.65
CdTe	II-VI	Ge_i	(+2/+1)	0.65	0.91
CdTe	II-VI	Ni_{Cd}	(0/-1)	0.92	0.56
CdTe	II-VI	Co_{Cd}	(0/-1)	1.25	0.94
CdTe	II-VI	Cr_{Cd}	(0/-1)	1.34	0.83
CdTe	II-VI	Sn_i	(+2/+1)	0.75	1.05
CdTe	II-VI	Pb_{Cd}	(+1/0)	0.22	0.17
CdTe	II-VI	P_{Te}	(0/-1)	0.07	0.04
CdSe	II-VI	Cu_{Cd}	(0/-1)	0.64	0.32
CdSe	II-VI	Cu_{Cd}	(-1/-2)	1.65	1.75
CdSe	II-VI	Ag_i	(+1/0)	0.99	1.20
CdSe	II-VI	Ag_i	(0/-1)	1.64	1.79
CdSe	II-VI	V_{Cd}	(-1/-2)	0.60	0.42
CdSe	II-VI	Cd_i	(+1/0)	1.65	1.41
CdSe	II-VI	V_{Se}	(+1/0)	1.52	1.11
CdSe	II-VI	V_{Se}	(+2/+1)	1.02	0.99
CdSe	II-VI	Se_i	(+1/0)	0.75	0.70
CdSe	II-VI	Fe_i	(+2/+1)	0.45	0.42
CdSe	II-VI	Co_i	(+1/0)	1.30	1.24
CdS	II-VI	V_{Cd}	(+1/0)	1.33	1.29
CdS	II-VI	Li_i	(+1/0)	1.61	1.59
CdS	II-VI	V_{Cd}	(-1/-2)	0.35	0.57
CdS	II-VI	V_S	(+2/0)	0.84	1.13
CdS	II-VI	Te_i	(+3/+2)	0.20	0.37
CdS	II-VI	I_S	(+1/0)	2.44	1.70
ZnTe	II-VI	N_{Te}	(0/-1)	0.05	0.00
ZnTe	II-VI	Cu_{Zn}	(0/-1)	0.15	0.23
ZnTe	II-VI	Cr_i	(+2/+1)	1.45	1.58
ZnSe	II-VI	Al_{Zn}	(+2/+1)	0.03	0.00
ZnSe	II-VI	F_i	(0/-1)	0.03	0.16
ZnSe	II-VI	Cl_i	(0/-1)	0.03	0.20
ZnSe	II-VI	I_i	(0/-1)	0.03	0.18
ZnSe	II-VI	Cs_{Zn}	(0/-1)	0.07	0.12
ZnSe	II-VI	V_{Zn}	(0/-1)	0.22	0.16
ZnS	II-VI	V_{Zn}	(-1/-2)	0.80	0.88
ZnS	II-VI	V_{Zn}	(0/-1)	0.50	0.27
ZnS	II-VI	V_S	(+2/0)	1.80	1.20
ZnO	II-VI	In_i	D, +3/+2	0.02	0.20
ZnO	II-VI	Cu_i	D, +2/+1	0.19	0.00
ZnO	II-VI	V_{Zn}	(0/-2)	0.07	0.00
ZnO	II-VI	O_i	(0/-1)	0.66	0.80
ZnO	II-VI	V_O	(+1/0)	2.00	1.86
ZnO	II-VI	V_O	(+2/+1)	0.32	0.62
ZnO	II-VI	Cu_i	(+1/0)	0.85	1.38
AlP	III-V	V_{Al}	(0/-1)	0.37	0.47
AlSb	III-V	V_{Al}	(0/-1)	0.04	0.06
GaN	III-V	V_N	(+2/+1)	0.11	0.00
GaN	III-V	V_Ga	(0/-1)	0.23	0.00
GaP	III-V	Ti_i	Acceptor, 0/-1	1.77	1.72

GaP	III-V	Ti_i	Donor, +1/0	1.00	1.31
GaP	III-V	Ni_{Ga}	Acceptor, 0/-1	0.51	0.39
GaP	III-V	Ni_{Ga}	Acceptor, -2/-3	1.45	1.62
GaAs	III-V	V_{Ga}	(-2/-3)	0.58	0.34
GaAs	III-V	As_{Ga}	(0/-1)	0.77	1.01
Si	IV-IV	Cu_i	Donor, +1/0	0.41	0.50
Si	IV-IV	Mn_i	Donor, +2/+1	0.26	0.26
Si	IV-IV	Mn_i	Donor, +1/0	0.76	0.77
Si	IV-IV	Mn_i	Acceptor, 0/-1	1.06	1.18
Si	IV-IV	O_{Si}	Donor, +1/0	1.12	1.06
Si	IV-IV	O_{Si}	Donor, +2/+1	1.05	0.71
Si	IV-IV	Tl_{Si}	Acceptor, 0/-1	0.25	0.36
Si	IV-IV	Ag_i	Donor, +1/0	0.45	0.57
Si	IV-IV	Ag_i	Acceptor, 0/-1	0.83	1.01
Si	IV-IV	Ge_{Si}	Acceptor, 0/-1	0.98	0.78
Si	IV-IV	C_i	Donor, +1/0	0.25	0.20
Si	IV-IV	W_i	Donor, +1/0	0.31	0.12
Si	IV-IV	W_i	Acceptor, 0/-1	0.75	0.72
Si	IV-IV	Sr	Donor, +2/+1	0.50	0.37
Si	IV-IV	Sr	Donor, +1/0	0.84	0.76
Ge	IV-IV	Au	(+1/0)	0.16	0.12
Ge	IV-IV	Au	(0/-1)	0.47	0.37
Ge	IV-IV	Au	(-1/-2)	0.63	0.58
Ge	IV-IV	Pt		0.04	0.00
SiC	IV-IV	V_{Si}	(0/-1)	0.50	0.76
SiC	IV-IV	V_C	(+1/0)	1.20	1.54
SiC	IV-IV	V_C	(+2/+1)	1.28	0.99
SiC	IV-IV	V_{Si}	(0/-1)	0.50	0.76

TABLE SIII: Dominant native defects in each compound and the equilibrium defect formation energy, Fermi level and conductivity determined by them, at A-rich chemical potential conditions.

Semiconductor	**Dominant Defects**	**Eqm. E^f (eV)**	**Eqm. E_F (eV, from VBM)**	**Eqm. Conductivity**
CdO	V_O, q = -1 and Cd_i, q = +1	1.16	0.45	moderately p-type
CdS	V_{Cd}, q = -2 and V_S, q = +1	1.82	1.52	moderately n-type
CdSe	V_{Cd}, q = -2 and Cd_i, q = +1	1.61	1.18	moderately n-type
CdTe	V_{Cd}, q = -2 and Cd_i, q = +1	1.48	0.70	intrinsic
ZnO	V_O, q = -1 and Zn_O, q = +2	1.94	0.89	moderately p-type
ZnS	V_{Zn}, q = -2 and Zn_S, q = +1	2.48	1.61	intrinsic
ZnSe	V_{Se}, q = -1 and Zn_{Se}, q = +1	2.25	1.38	intrinsic
ZnTe	V_{Zn}, q = -2 and Zn_i, q = +1	1.58	0.83	moderately p-type
BN	V_B, q = -1 and V_N, q = +1	5.29	3.70	moderately n-type
BP	B_P, q = -1 and P_B, q = +2	2.57	-0.09	very p-type
BAs	V_B, q = +1 and B_{As}, q = -1	2.33	-0.13	very p-type
BSb	V_B, q = -1 and Sb_B, q = +3	1.30	-0.15	very p-type
AlN	V_{Al}, q = -2 and V_N, q = +1	4.15	4.44	moderately n-type
AlP	Al_P, q = -1 and Al_i, q = +2	3.27	1.12	intrinsic
AlAs	Al_{As}, q = -1 and Al_i, q = +2	2.80	0.74	moderately p-type
AlSb	V_{Al}, q = -3 and Sb_i, q = +3	1.18	0.97	moderately n-type
GaN	V_{Ga}, q = -3 and V_N, q = +1	3.82	3.07	very n-type
GaP	V_P, q = +1 and Ga_P, q = -1	2.16	0.26	moderately p-type
GaAs	Ga_{As}, q = -1 and Ga_i, q = +1	1.62	-0.16	very p-type
GaSb	Ga_{Sb}, q = -1 and Ga_i, q = +1	0.81	-0.15	very p-type
InN	V_N, q = -1 and In_i, q = +2	2.47	0.09	moderately p-type
InP	V_P, q = +1 and In_P, q = -1	2.14	0.65	intrinsic
InAs	V_{As}, q = -2 and In_i, q = +1	1.90	0.58	very n-type
InSb	In_{Sb}, q = -1 and In_i, q = +1	1.01	0.22	moderately p-type
C	V_C, q = -1 and C_i, q = +3	11.06	-2.60	very p-type
Si	V_{Si}, q = +1 and Si_i, q = -1	4.11	0.61	intrinsic
Ge	V_{Ge}, q = -1 and Ge_i, q = +3	2.21	-0.26	very p-type
Sn	V_{Sn}, q = -2 and Sn_i, q = +2	1.3	-0.11	very p-type
SiC	V_C, q = +2 and C_{Si}, q = -1	4.54	1.30	intrinsic
GeC	V_C, q = +1 and C_{Ge}, q = -1	3.36	0.69	intrinsic
SnC	V_C, q = +2 and C_{Sn}, q = -1	2.96	0.60	intrinsic
SiGe	Ge_{Si}, q = +1 and Si_{Ge}, q = -1	0.42	0.51	moderately n-type
SiSn	Sn_{Si}, q = +1 and Si_{Sn}, q = -1	0.29	0.75	very n-type
GeSn	Sn_{Ge}, q = +1 and Ge_{Sn}, q = -1	0.19	0.15	moderately p-type

TABLE SIV: Dominant native defects in each compound and the equilibrium defect formation energy, Fermi level and conductivity determined by them, at B-rich chemical potential conditions.

Semiconductor	**Dominant Defects**	**Eqm. E^f (eV)**	**Eqm. E_F (eV, from VBM)**	**Eqm. Conductivity**
CdO	V_{Cd}, q = -2 and V_O, q = +1	2.72	0.75	moderately p-type
CdS	V_{Cd}, q = -2 and V_S, q = +2	2.02	0.77	moderately p-type
CdSe	V_{Cd}, q = -2 and Cd_i, q = +2	1.73	0.48	moderately p-type
CdTe	V_{Cd}, q = -1 and Te_{Cd}, q = +1	1.54	0.14	very p-type
ZnO	V_{Zn}, q = -1 and O_i, q = +1	3.41	-0.06	very p-type
ZnS	V_{Zn}, q = -1 and Zn_i, q = +2	2.39	0.43	moderately p-type
ZnSe	V_{Zn}, q = -1 and Se_{Zn}, q = +2	1.85	0.19	very p-type
ZnTe	V_{Zn}, q = -1 and Zn_i, q = +2	1.49	-0.01	very p-type
BN	N_B, q = +1 and N_i, q = -1	0.99	5.66	very n-type
BP	B_P, q = -1 and P_B, q = +1	2.88	1.23	moderately n-type
BAs	B_{As}, q = -1 and As_B, q = +2	2.29	0.06	very p-type
BSb	V_B, q = +1 and B_{Sb}, q = -1	1.54	-0.36	very p-type
AlN	N_{Al}, q = -1 and N_i, q = +1	-1.46	0.40	very p-type
AlP	V_{Al}, q = -2 and P_{Al}, q = +2	2.78	1.29	intrinsic
AlAs	V_{Al}, q = -2 and As_{Al}, q = +1	1.73	1.17	intrinsic
AlSb	V_{Al}, q = -3 and Sb_i, q = +3	0.82	0.97	moderately n-type
GaN	V_{Ga}, q = -2 and N_{Ga}, q = +1	-1.00	2.38	moderately n-type
GaP	V_{Ga}, q = -3 and P_{Ga}, q = +1	2.20	1.08	intrinsic
GaAs	V_{Ga}, q = -2 and As_{Ga}, q = +2	1.73	0.32	moderately p-type
GaSb	Ga_{Sb}, q = -2 and Sb_{Ga}, q = +1	1.12	0.21	moderately p-type
InN	V_{In}, q = -3 and N_{In}, q = +1	-1.15	1.93	very n-type
InP	V_P, q = -2 and P_{In}, q = +1	2.26	1.04	moderately n-type
InAs	V_{In}, q = -2 and As_{In}, q = +1	1.52	0.67	very n-type
InSb	V_{In}, q = -1 and Sb_{In}, q = +1	1.32	0.56	very n-type
C	V_C, q = -1 and C_i, q = +1	12.95	-4.48	very p-type
Si	V_{Si}, q = +1 and Si_i, q = -1	4.11	0.61	intrinsic
Ge	V_{Ge}, q = -1 and Ge_i, q = +2	2.17	-0.21	very p-type
Sn	V_{Sn}, q = -1 and Sn_i, q = +2	1.35	-0.09	very p-type
SiC	V_C, q = +1 and C_{Si}, q = -1	4.17	0.83	moderately p-type
GeC	V_C, q = -1 and Ge_C, q = +1	2.90	1.33	very n-type
SnC	V_C, q = -1 and Sn_C, q = +2	1.55	0.41	moderately p-type
SiGe	Ge_{Si}, q = +1 and Si_{Ge}, q = -1	0.42	0.32	moderately p-type
SiSn	Sn_{Si}, q = +1 and Si_{Sn}, q = -1	0.29	0.07	moderately p-type
GeSn	Sn_{Ge}, q = +1 and Ge_{Sn}, q = -1	0.19	0.04	very p-type

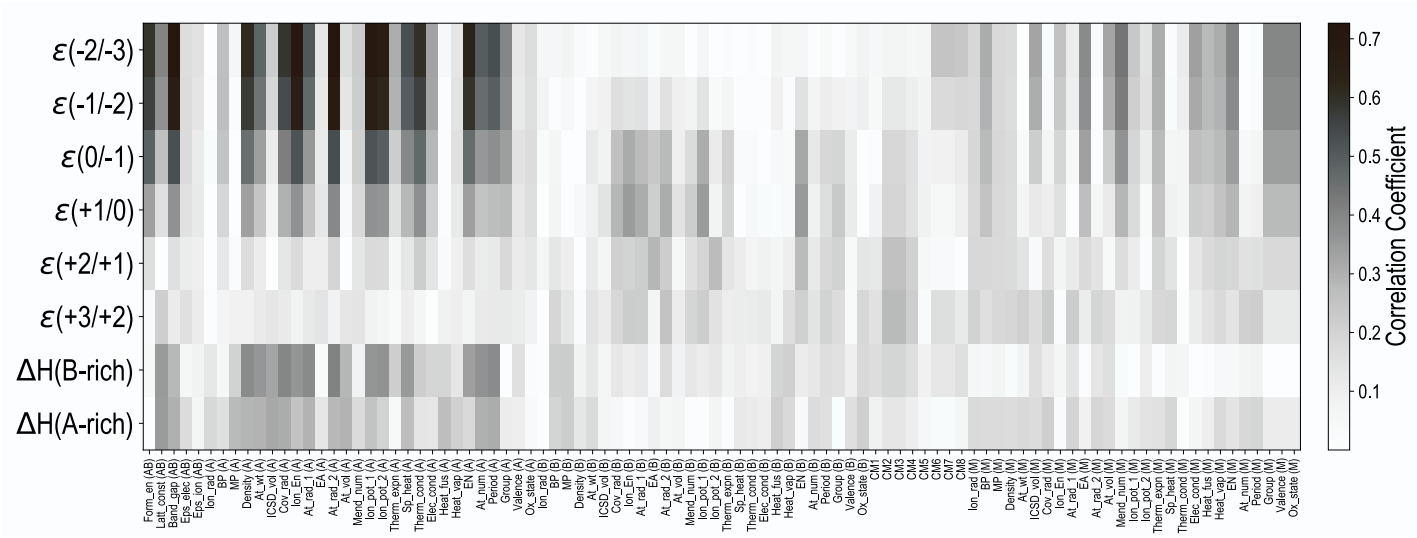

FIG. S2: Absolute values of coefficient of linear correlation between every unique descriptor and every property; darker boxes imply high correlation whereas white boxes mean there is zero correlation. The descriptor labels and the corresponding linear correlation values can also be found in Table SV.

TABLE SV: Tabulated absolute values of coefficient of linear correlation between every unique descriptor and every property.

Labels	**ΔH (A-rich)**	**ΔH (B-rich)**	**ε(+3/+2)**	**ε(+2/+1)**	**ε(+1/0)**	**ε(0/-1)**	**ε(-1/-2)**	**ε(-2/-3)**
Form en (AB)	0.01	0.05	0.04	0.17	0.34	0.49	0.56	0.59
Latt const (AB)	0.35	0.35	0.22	0.00	0.15	0.26	0.37	0.40
Band gap (AB)	0.30	0.28	0.12	0.16	0.37	0.53	0.66	0.72
Eps elec (AB)	0.17	0.06	0.16	0.11	0.15	0.17	0.17	0.17
Eps ion (AB)	0.07	0.09	0.11	0.10	0.12	0.12	0.14	0.15
Ion rad (A)	0.19	0.13	0.02	0.00	0.00	0.03	0.02	0.01
BP (A)	0.16	0.06	0.08	0.13	0.19	0.25	0.27	0.26
MP (A)	0.28	0.21	0.09	0.03	0.01	0.01	0.03	0.06
Density (A)	0.29	0.38	0.09	0.17	0.34	0.46	0.57	0.61
At wt (A)	0.29	0.36	0.12	0.11	0.25	0.34	0.44	0.48
ICSD vol (A)	0.32	0.33	0.14	0.00	0.07	0.12	0.19	0.22
Cov rad (A)	0.31	0.40	0.12	0.13	0.30	0.42	0.54	0.59
Ion En (A)	0.25	0.36	0.09	0.18	0.37	0.52	0.65	0.69
At rad 1 (A)	0.29	0.39	0.12	0.10	0.25	0.36	0.47	0.51
EA (A)	0.13	0.04	0.14	0.10	0.13	0.12	0.12	0.13
At rad 2 (A)	0.28	0.41	0.08	0.19	0.39	0.54	0.68	0.73
At vol (A)	0.29	0.29	0.13	0.01	0.04	0.07	0.14	0.17
Mend num (A)	0.17	0.08	0.14	0.14	0.18	0.22	0.22	0.23
Ion pot 1 (A)	0.25	0.36	0.09	0.18	0.37	0.52	0.65	0.69
Ion pot 2 (A)	0.21	0.37	0.07	0.18	0.36	0.49	0.62	0.67
Therm expn (A)	0.04	0.22	0.04	0.06	0.16	0.22	0.30	0.31
Sp heat (A)	0.27	0.35	0.10	0.13	0.28	0.40	0.50	0.53
Therm cond (A)	0.14	0.22	0.06	0.22	0.38	0.47	0.56	0.60
Elec cond (A)	0.13	0.19	0.02	0.14	0.22	0.29	0.32	0.33
Heat fus (A)	0.27	0.20	0.09	0.04	0.02	0.02	0.02	0.05
Heat vap (A)	0.21	0.13	0.10	0.10	0.12	0.16	0.15	0.13
EN (A)	0.18	0.34	0.07	0.16	0.34	0.46	0.59	0.63
At num (A)	0.30	0.37	0.12	0.11	0.25	0.36	0.46	0.50
Period (A)	0.31	0.38	0.12	0.12	0.27	0.38	0.49	0.53
Group (A)	0.11	0.01	0.12	0.18	0.28	0.34	0.38	0.40
Valence (A)	0.10	0.15	0.11	0.12	0.13	0.15	0.12	0.13
Ox state (A)	0.04	0.03	0.08	0.13	0.17	0.17	0.16	0.17
Ion rad (B)	0.01	0.02	0.01	0.06	0.01	0.03	0.08	0.06
BP (B)	0.20	0.22	0.19	0.13	0.07	0.07	0.02	0.06
MP (B)	0.22	0.23	0.14	0.10	0.01	0.01	0.05	0.07
Density (B)	0.14	0.09	0.04	0.00	0.05	0.01	0.01	0.03
At wt (B)	0.17	0.13	0.15	0.12	0.09	0.12	0.07	0.02
ICSD vol (B)	0.06	0.07	0.11	0.07	0.15	0.09	0.09	0.05
Cov rad (B)	0.05	0.13	0.19	0.21	0.29	0.26	0.16	0.07
Ion En (B)	0.02	0.11	0.22	0.20	0.35	0.31	0.15	0.06
At rad 1 (B)	0.04	0.12	0.22	0.22	0.31	0.27	0.15	0.07
EA (B)	0.10	0.04	0.07	0.29	0.22	0.24	0.08	0.03
At rad 2 (B)	0.06	0.17	0.25	0.22	0.31	0.28	0.15	0.05
At vol (B)	0.07	0.08	0.11	0.05	0.15	0.10	0.10	0.06
Mend num (B)	0.02	0.13	0.19	0.16	0.22	0.18	0.08	0.01
Ion pot 1 (B)	0.01	0.11	0.22	0.19	0.35	0.31	0.15	0.06
Ion pot 2 (B)	0.07	0.13	0.15	0.27	0.08	0.11	0.04	0.01
Therm expn (B)	0.02	0.09	0.11	0.14	0.20	0.21	0.07	0.02
Sp heat (B)	0.12	0.10	0.11	0.11	0.01	0.04	0.02	0.00
Therm cond (B)	0.11	0.08	0.12	0.07	0.05	0.04	0.02	0.02
Elec cond (B)	0.10	0.09	0.12	0.09	0.03	0.01	0.02	0.01
Heat fus (B)	0.22	0.20	0.08	0.06	0.03	0.05	0.10	0.08
Heat vap (B)	0.18	0.21	0.16	0.10	0.04	0.04	0.03	0.06
EN (B)	0.05	0.11	0.21	0.28	0.35	0.32	0.15	0.06
At num (B)	0.17	0.13	0.15	0.12	0.09	0.12	0.07	0.02
Period (B)	0.16	0.16	0.20	0.19	0.17	0.19	0.11	0.04
Group (B)	0.03	0.15	0.17	0.21	0.23	0.20	0.09	0.01
Valence (B)	0.16	0.10	0.04	0.04	0.11	0.11	0.08	0.04

Ox state (B)	0.21	0.18	0.19	0.13	0.01	0.01	0.10	0.08
CM1	0.09	0.17	0.15	0.13	0.08	0.02	0.04	0.05
CM2	0.13	0.20	0.28	0.26	0.19	0.19	0.12	0.07
CM3	0.13	0.20	0.28	0.26	0.19	0.19	0.12	0.07
CM4	0.10	0.17	0.22	0.21	0.15	0.16	0.10	0.06
CM5	0.05	0.09	0.03	0.05	0.09	0.08	0.08	0.06
CM6	0.03	0.13	0.17	0.04	0.05	0.09	0.18	0.25
CM7	0.03	0.13	0.17	0.04	0.05	0.09	0.18	0.25
CM8	0.06	0.16	0.12	0.01	0.10	0.12	0.19	0.24
Ion rad (M)	0.18	0.05	0.21	0.19	0.18	0.19	0.18	0.19
BP (M)	0.18	0.05	0.16	0.19	0.25	0.28	0.30	0.31
MP (M)	0.18	0.07	0.18	0.18	0.18	0.19	0.18	0.18
Density (M)	0.17	0.03	0.19	0.18	0.17	0.18	0.17	0.17
At wt (M)	0.18	0.07	0.21	0.15	0.07	0.04	0.00	0.01
ICSD vol (M)	0.17	0.16	0.17	0.04	0.12	0.22	0.31	0.33
Cov rad (M)	0.21	0.11	0.23	0.16	0.11	0.07	0.04	0.03
Ion En (M)	0.04	0.07	0.03	0.07	0.16	0.20	0.23	0.26
At rad 1 (M)	0.20	0.14	0.22	0.12	0.02	0.05	0.10	0.11
EA (M)	0.09	0.02	0.11	0.18	0.27	0.34	0.38	0.39
At rad 2 (M)	0.19	0.14	0.20	0.14	0.06	0.03	0.02	0.03
At vol (M)	0.17	0.16	0.17	0.04	0.12	0.22	0.31	0.32
Mend num (M)	0.08	0.03	0.09	0.17	0.29	0.37	0.42	0.44
Ion pot 1 (M)	0.12	0.01	0.09	0.11	0.17	0.19	0.21	0.24
Ion pot 2 (M)	0.14	0.10	0.13	0.03	0.04	0.11	0.15	0.15
Therm expn (M)	0.15	0.02	0.19	0.20	0.25	0.28	0.30	0.31
Sp heat (M)	0.19	0.07	0.19	0.15	0.09	0.06	0.03	0.03
Therm cond (M)	0.02	0.06	0.02	0.01	0.07	0.07	0.10	0.12
Elec cond (M)	0.07	0.00	0.11	0.13	0.22	0.26	0.32	0.34
Heat fus (M)	0.10	0.02	0.15	0.17	0.21	0.25	0.28	0.28
Heat vap (M)	0.18	0.05	0.18	0.20	0.25	0.28	0.30	0.31
EN (M)	0.14	0.04	0.14	0.19	0.29	0.35	0.39	0.40
At num (M)	0.19	0.08	0.21	0.15	0.08	0.05	0.01	0.00
Period (M)	0.20	0.09	0.22	0.16	0.10	0.07	0.03	0.03
Group (M)	0.11	0.01	0.12	0.18	0.28	0.34	0.38	0.40
Valence (M)	0.11	0.01	0.12	0.18	0.28	0.34	0.38	0.40
Ox state (M)	0.11	0.01	0.12	0.18	0.28	0.34	0.38	0.40

TABLE SVI: ML training 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)**
ε(+3/+2)	MLR	0.316	0.313	0.283	0.308
ε(+3/+2)	Ridge	0.324	0.319	0.284	0.314
ε(+3/+2)	LASSO	0.331	0.326	0.287	0.320
ε(+3/+2)	Elastic Net	0.323	0.319	0.284	0.313
ε(+3/+2)	RFR	0.213	0.255	0.203	0.222
ε(+3/+2)	KRR	0.231	0.259	0.200	0.231
ε(+3/+2)	GPR	0.237	0.262	0.209	0.237
ε(+3/+2)	NN	0.189	0.227	0.160	0.193
ε(+2/+1)	MLR	0.390	0.383	0.418	0.395
ε(+2/+1)	Ridge	0.403	0.389	0.422	0.404
ε(+2/+1)	LASSO	0.408	0.390	0.422	0.407
ε(+2/+1)	Elastic Net	0.401	0.387	0.420	0.403
ε(+2/+1)	RFR	0.243	0.290	0.275	0.262
ε(+2/+1)	KRR	0.218	0.217	0.231	0.221
ε(+2/+1)	GPR	0.177	0.172	0.178	0.176
ε(+2/+1)	NN	0.202	0.201	0.208	0.203
ε(+1/0)	MLR	0.371	0.329	0.377	0.363
ε(+1/0)	Ridge	0.380	0.343	0.390	0.374
ε(+1/0)	LASSO	0.386	0.347	0.393	0.379
ε(+1/0)	Elastic Net	0.378	0.342	0.389	0.373
ε(+1/0)	RFR	0.248	0.278	0.257	0.257
ε(+1/0)	KRR	0.191	0.168	0.202	0.189
ε(+1/0)	GPR	0.122	0.101	0.127	0.119
ε(+1/0)	NN	0.201	0.156	0.211	0.194
ε(0/-1)	MLR	0.333	0.324	0.279	0.319
ε(0/-1)	Ridge	0.342	0.337	0.303	0.332
ε(0/-1)	LASSO	0.341	0.338	0.301	0.331
ε(0/-1)	Elastic Net	0.341	0.336	0.302	0.331
ε(0/-1)	RFR	0.213	0.236	0.205	0.217
ε(0/-1)	KRR	0.205	0.179	0.171	0.192
ε(0/-1)	GPR	0.129	0.089	0.099	0.114
ε(0/-1)	NN	0.204	0.167	0.190	0.193
ε(-1/-2)	MLR	0.294	0.286	0.253	0.283
ε(-1/-2)	Ridge	0.301	0.306	0.258	0.293
ε(-1/-2)	LASSO	0.301	0.305	0.257	0.292
ε(-1/-2)	Elastic Net	0.299	0.300	0.255	0.289
ε(-1/-2)	RFR	0.221	0.262	0.216	0.230
ε(-1/-2)	KRR	0.178	0.157	0.145	0.166
ε(-1/-2)	GPR	0.134	0.106	0.101	0.121
ε(-1/-2)	NN	0.208	0.176	0.200	0.199
ε(-2/-3)	MLR	0.253	0.212	0.194	0.231
ε(-2/-3)	Ridge	0.257	0.224	0.198	0.237
ε(-2/-3)	LASSO	0.257	0.223	0.198	0.236
ε(-2/-3)	Elastic Net	0.257	0.223	0.197	0.236
ε(-2/-3)	RFR	0.202	0.182	0.167	0.190
ε(-2/-3)	KRR	0.232	0.161	0.156	0.201
ε(-2/-3)	GPR	0.182	0.133	0.130	0.160
ε(-2/-3)	NN	0.215	0.142	0.169	0.190

TABLE SVII: ML training 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	0.786	1.265	1.439	0.982
ΔH (A-rich)	Ridge	0.796	1.273	1.460	0.993
ΔH (A-rich)	LASSO	0.832	1.322	1.505	1.033
ΔH (A-rich)	Elastic Net	0.800	1.279	1.466	0.998
ΔH (A-rich)	RFR	0.440	0.860	1.045	0.632
ΔH (A-rich)	KRR	0.436	0.678	0.657	0.513
ΔH (A-rich)	GPR	0.318	0.482	0.519	0.380
ΔH (A-rich)	NN	0.489	0.645	0.765	0.560
ΔH (B-rich)	MLR	0.960	1.330	1.387	1.089
ΔH (B-rich)	Ridge	0.974	1.345	1.419	1.106
ΔH (B-rich)	LASSO	1.015	1.397	1.495	1.155
ΔH (B-rich)	Elastic Net	0.976	1.347	1.424	1.108
ΔH (B-rich)	RFR	0.556	0.918	1.074	0.711
ΔH (B-rich)	KRR	0.504	0.680	0.666	0.558
ΔH (B-rich)	GPR	0.527	0.773	0.793	0.611
ΔH (B-rich)	NN	0.634	0.713	0.723	0.659

Description of balanced vs unbalanced dataset ML comparison: In the original dataset, there is 70% II-VI, 15% III-V and 15% IV-IV data for the semiconductor type. And there is 60% interstitial site defect but only 40% substitutional defect. To balance the type of dataset, we randomly removed the data point of CdTe, as it takes up the majority of II-VI. For the site balance, we randomly dropped out interstitial site defects to match the quantity with subsitional defects. Finally, we obtained one dataset with equal amounts of each semiconductor type, the other one with equivalent interstitial site and substitutional defects. For KRR, GPR and all the linear models, the RMSEs became worse or comparable to the non-balanced dataset. While the balanced dataset did narrow the gap between training and testing RMSE. However, even using the balanced dataset, the RMSE for II-IV is still better than III-V and IV-IV and the RMSE for interstitial site defect is better than the substitutional defects. For the RFR Model, we found that model performance suffered significantly when tested on the balanced datasets. Model prediction on the test set became worse for each of the 8 targets, with larger prediction losses using the dataset balanced by type of semiconductor vs the dataset balanced by impurity substitution site. For both datasets, the formation energies showed the largest decrease in test set prediction. In all cases the overfitting also became worse, and followed trends similar to the test set prediction where the largest increase in overfitting was found using the dataset balanced by type of semiconductor, and overfitting of the formation energy targets was more extreme than the transition energy level targets.

As a further test of overfitting tendency, we compared models trained using only the original features to those using SISSO-based features. In the RFR model, comparing predictions from data sets using the original descriptors (compared with the SISSO expanded features) found that the models predicted slightly worse for each of the eight targets. But, as the reviewer mentioned, using the original descriptors mitigated overfitting of the training, most obviously in prediction of the formation energies. For dHA and dHB, the improvement in overfitting on the train set was larger than the improvement in prediction of the test set. For the transition energy levels, there was very little difference in overfitting comparing the original descriptors to the SISSO expanded descriptors. For KRR and GPR models, SISSO features do not improve the model performance. But for all the linear models, SISSO features help them quite a lot. The reason could be that both KRR and GPR are non-linear models, they are able to internally extract non-linear relations between descriptors while other linear models could not. We notice that using the original features instead of SISSO features leads to lesser overfitting (smaller difference between training and test errors), but yields similar test errors; thus, we ultimately report the SISSO-based models for each ML technique.

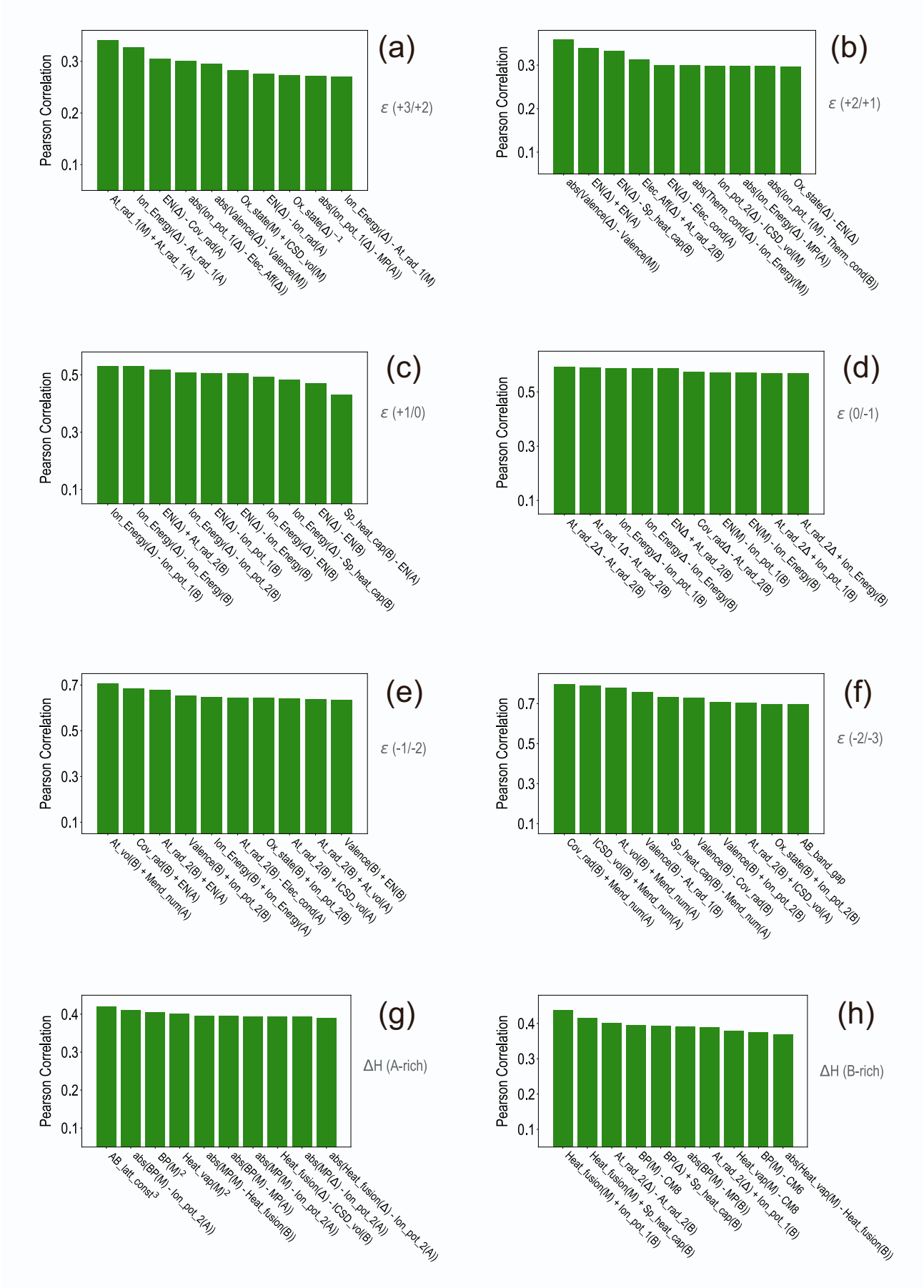

FIG. S3: Linear correlation coefficients for the ten best SISSO-based compound features, for all 8 properties. Δ refers to the difference of the corresponding property (feature) between the impurity atom M and the element previously occupying the site it exists in (A, B or none).

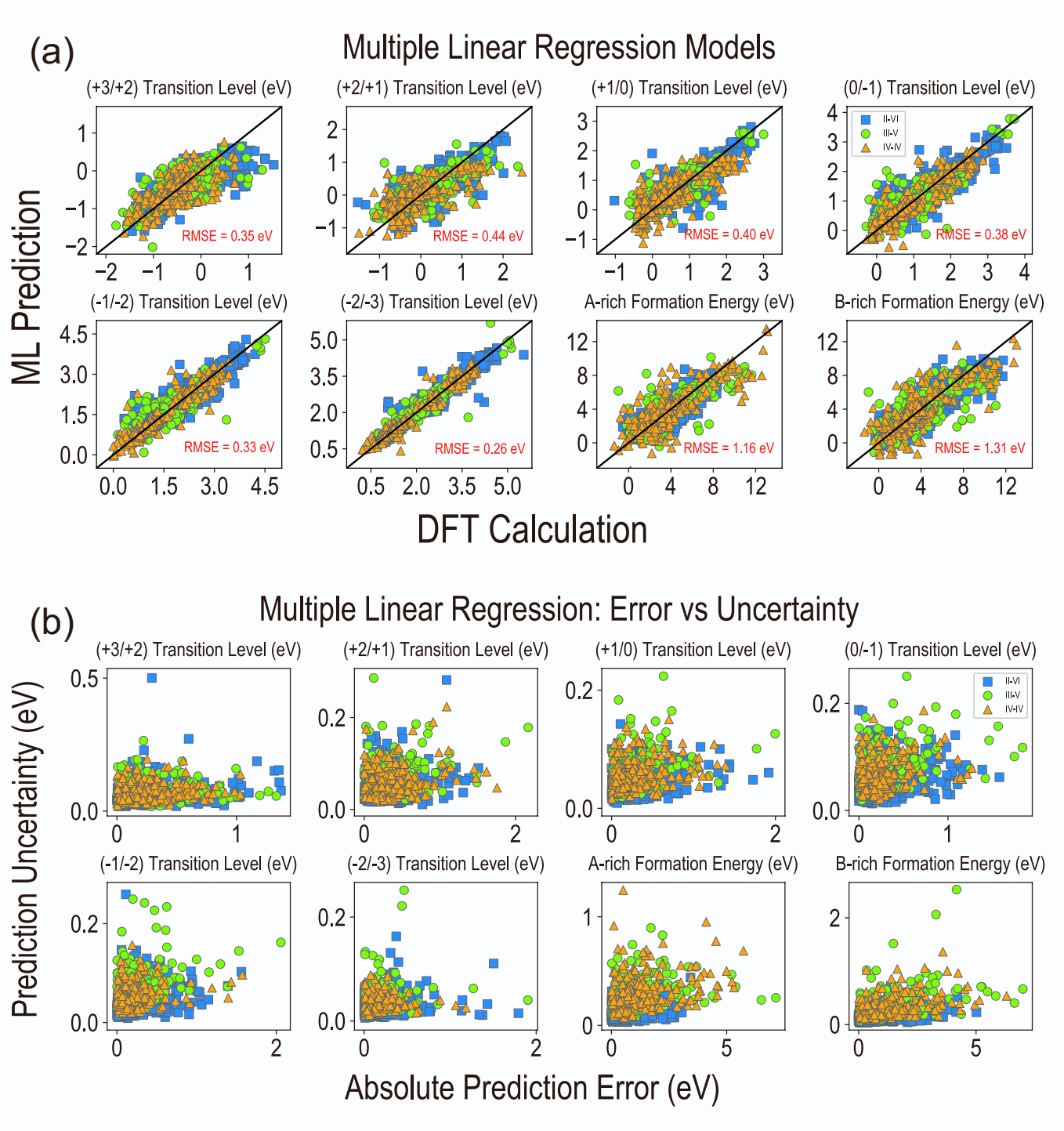

FIG. S4: Multiple Linear Regression results: (a) parity plots, and (b) prediction uncertainty as a function of absolute prediction error.

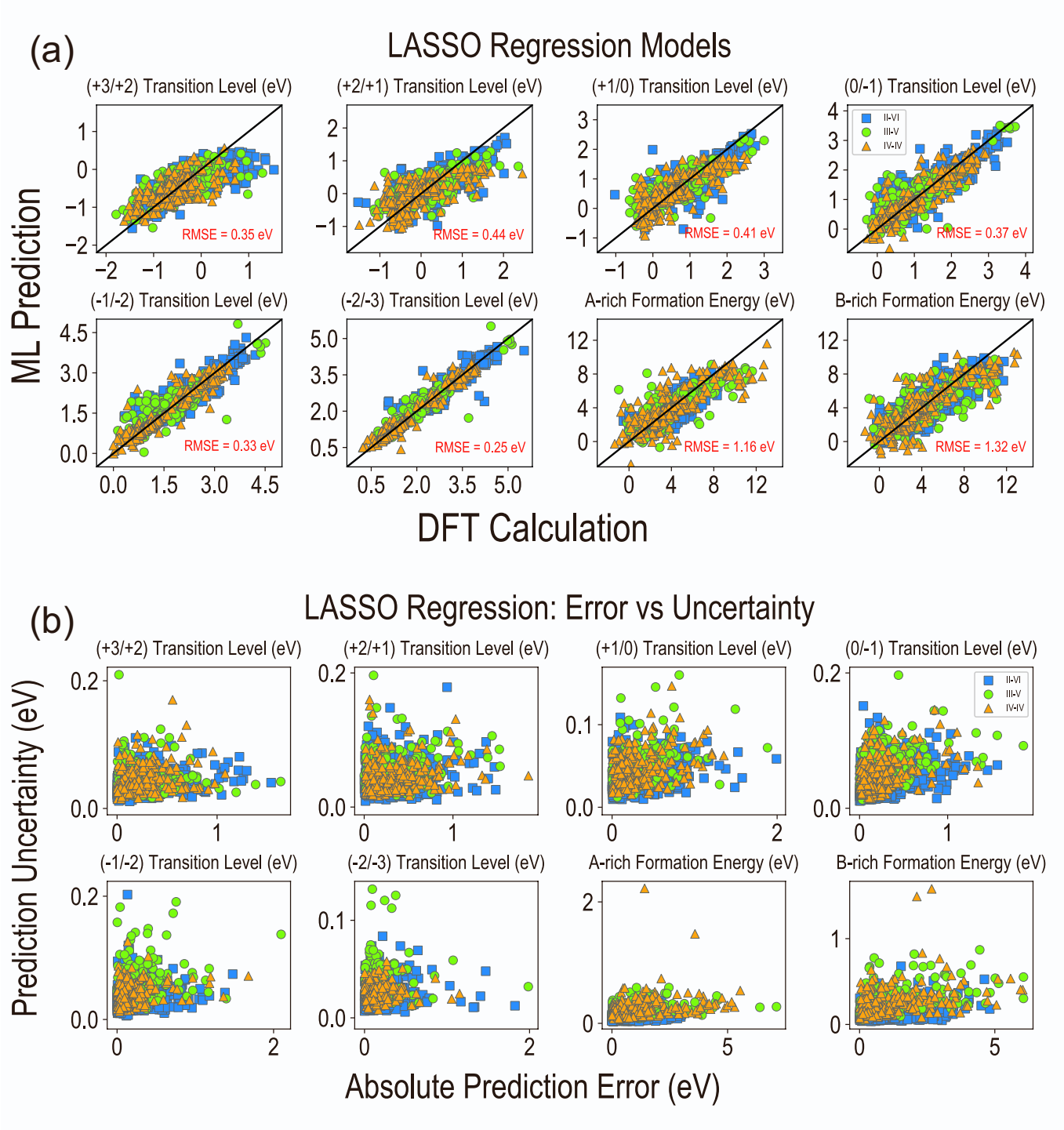

FIG. S5: LASSO Regression results: (a) parity plots, and (b) prediction uncertainty as a function of absolute prediction error.

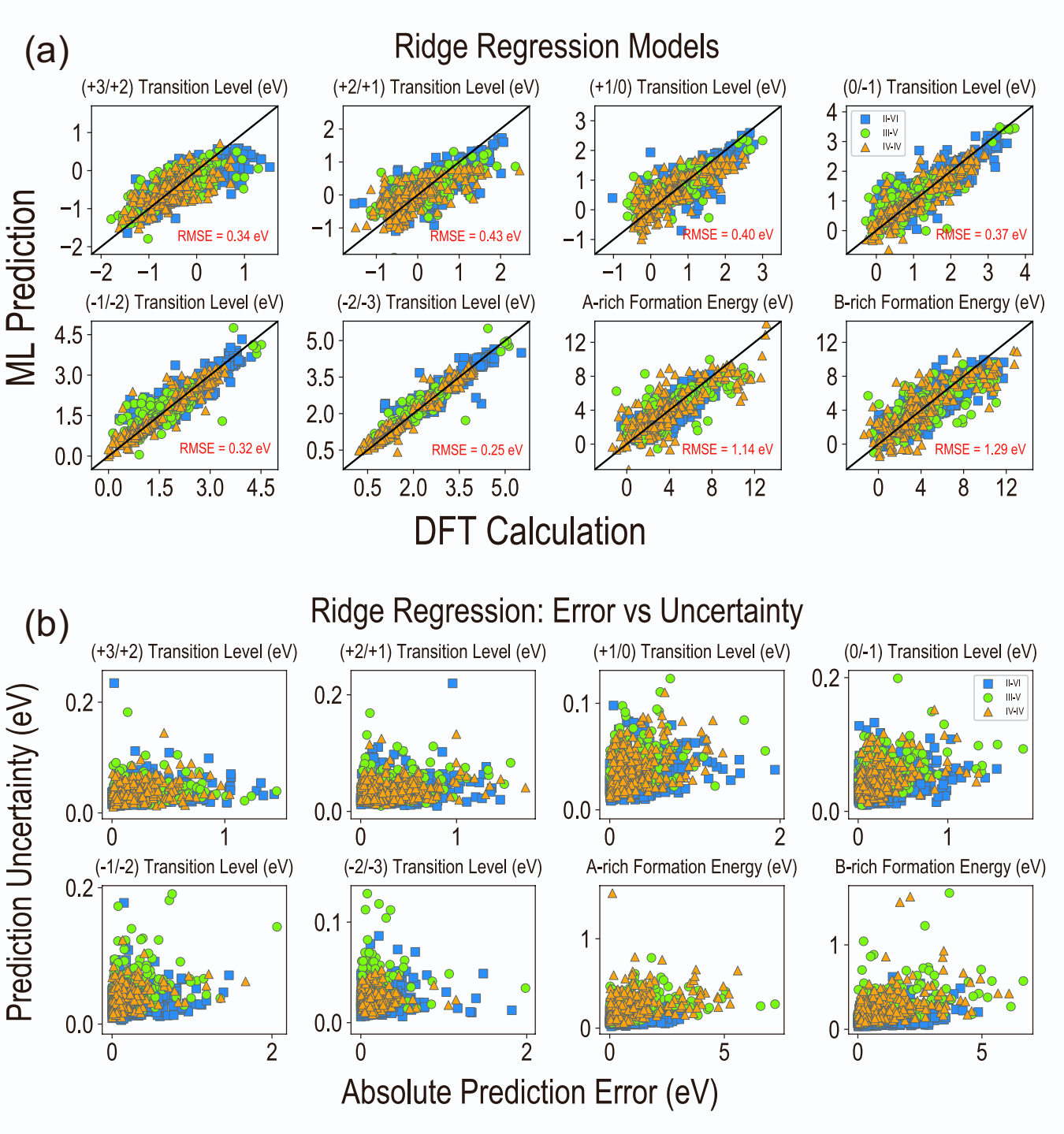

FIG. S6: Ridge Regression results: (a) parity plots, and (b) prediction uncertainty as a function of absolute prediction error.

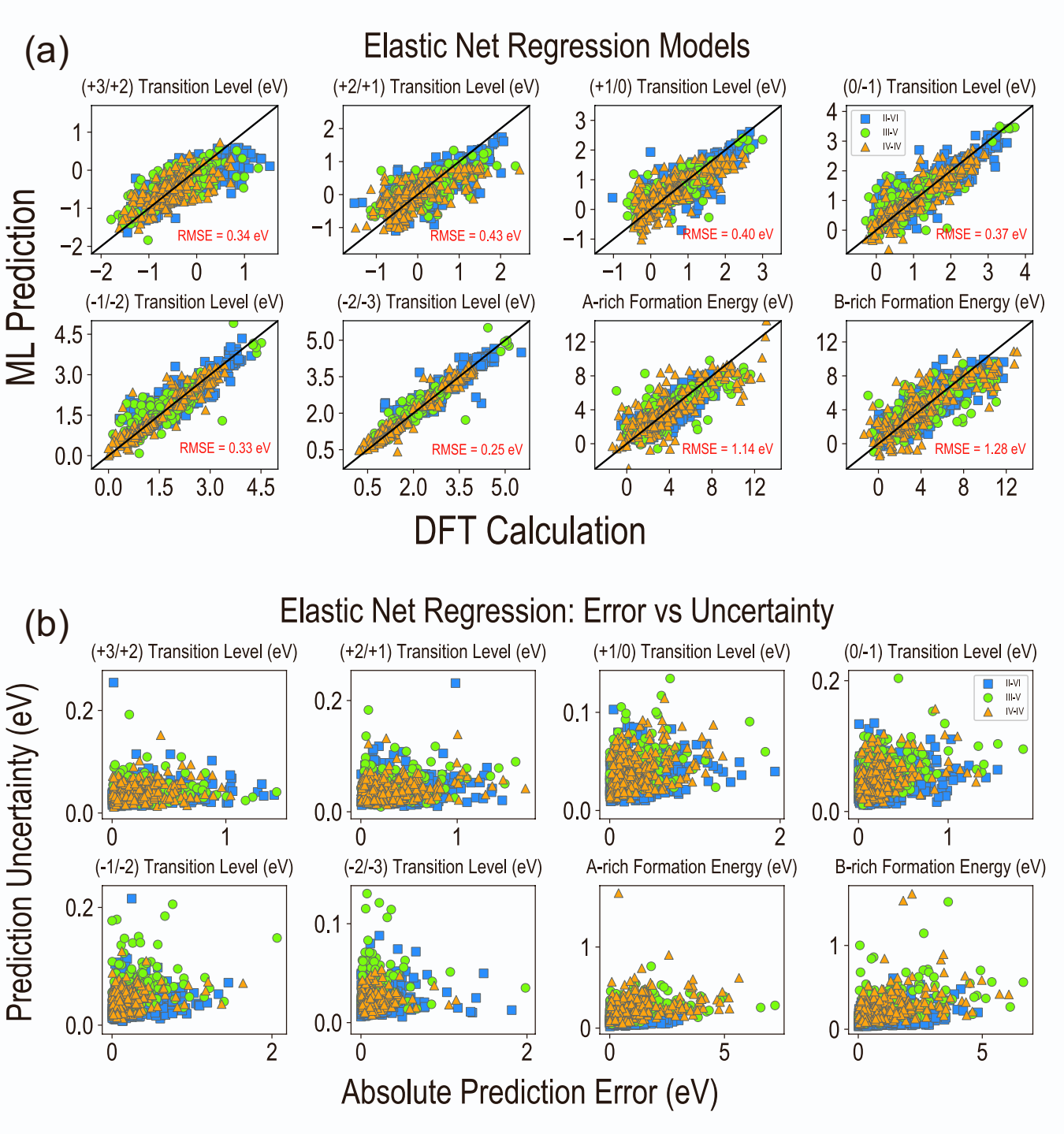

FIG. S7: Elastic Net Regression results: (a) parity plots, and (b) prediction uncertainty as a function of absolute prediction error.

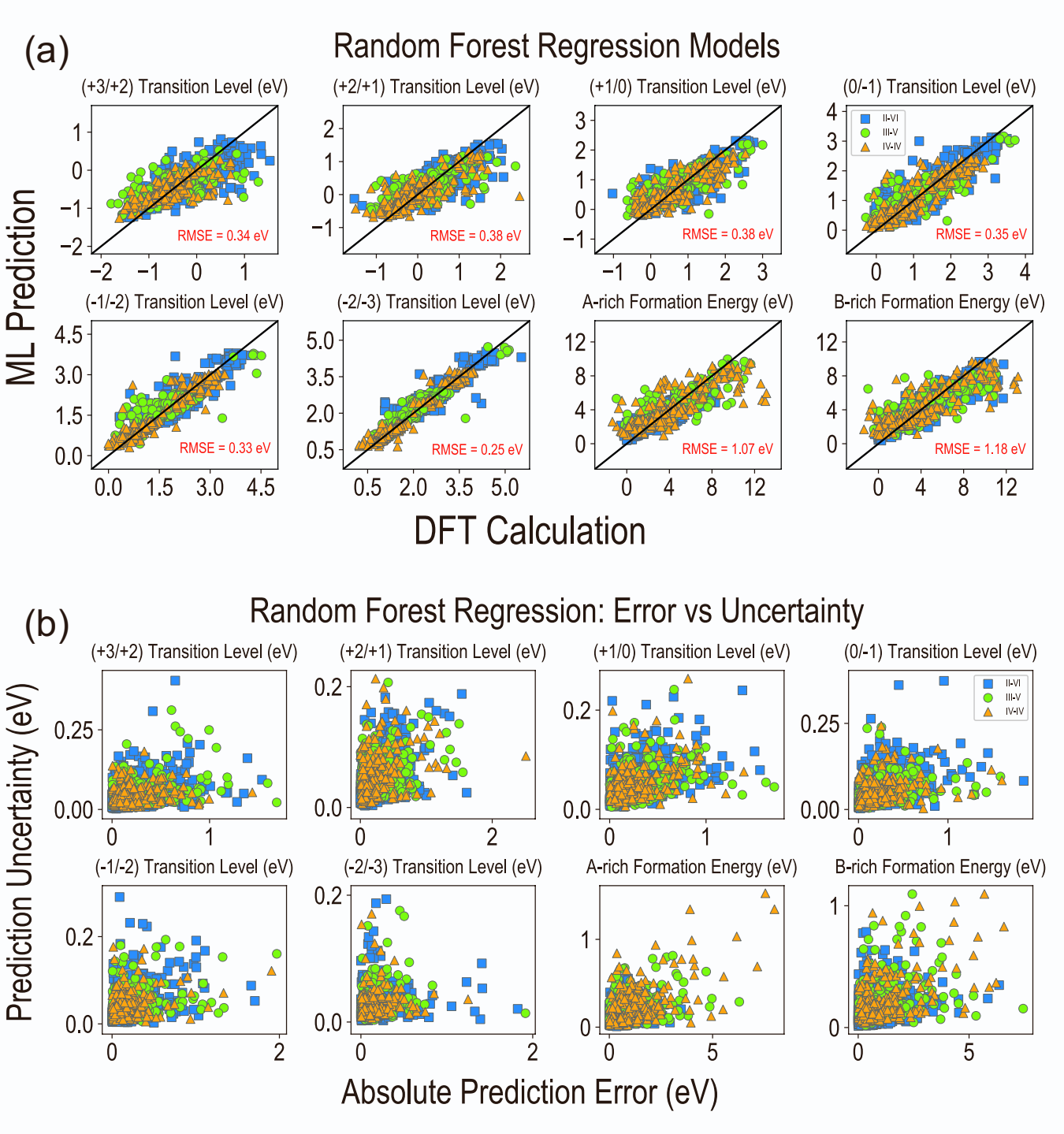

FIG. S8: Random Forest Regression (RFR) results: (a) parity plots, and (b) prediction uncertainty as a function of absolute prediction error.

SISSO/LASSO DATASET - Bayesian Opt.		
SISSO/LASSO manuscript predictions	**train RMSE**	**test RMSE**
dHA	0.632	1.07
dHB	0.711	1.18
(+3/+2)	0.222	0.34
(+2/+1)	0.262	0.38
(+1/0)	0.257	0.38
(0/-1)	0.217	0.35
(-1/-2)	0.23	0.33
(-2/-3)	0.19	0.25

NEW SISSO/LASSO DATASET - Grid Search		
SISSO/LASSO Grid Search	**train RMSE**	**test RMSE**
dHA	0.42	1.02
dHB	0.5	1.13
(+3/+2)	0.18	0.35
(+2/+1)	0.17	0.37
(+1/0)	0.15	0.36
(0/-1)	0.14	0.35
(-1/-2)	0.15	0.32
(-2/-3)	0.14	0.25

FIG. S9: Comparison of RFR performance with hyperparameter optimization based on grid search and Bayesian search: for the RFR model, we looked into GridSearch for hyperparameter optimization to compare against the Bayesian method used in the manuscript. From the sklearn model-selection library, we employed RandomizedSearchCV to investigate a large hyperparameter space, followed by GridSearchCV to more finely tune hyperparameters based on best results from the random search. We found that for the formation enthalpies the test RMSE slightly improved at the expense of large model overfitting, potentially because it is hard to mitigate overfitting using the GridSearch functionality, unlike Bayesian where loss could be adjusted to steer away from overfitting. For transition energy level predictions, the model predictions were nearly the same or slightly worse and overfitting again became more extreme. We thus conclude that Bayesian search serves us better for RFR.

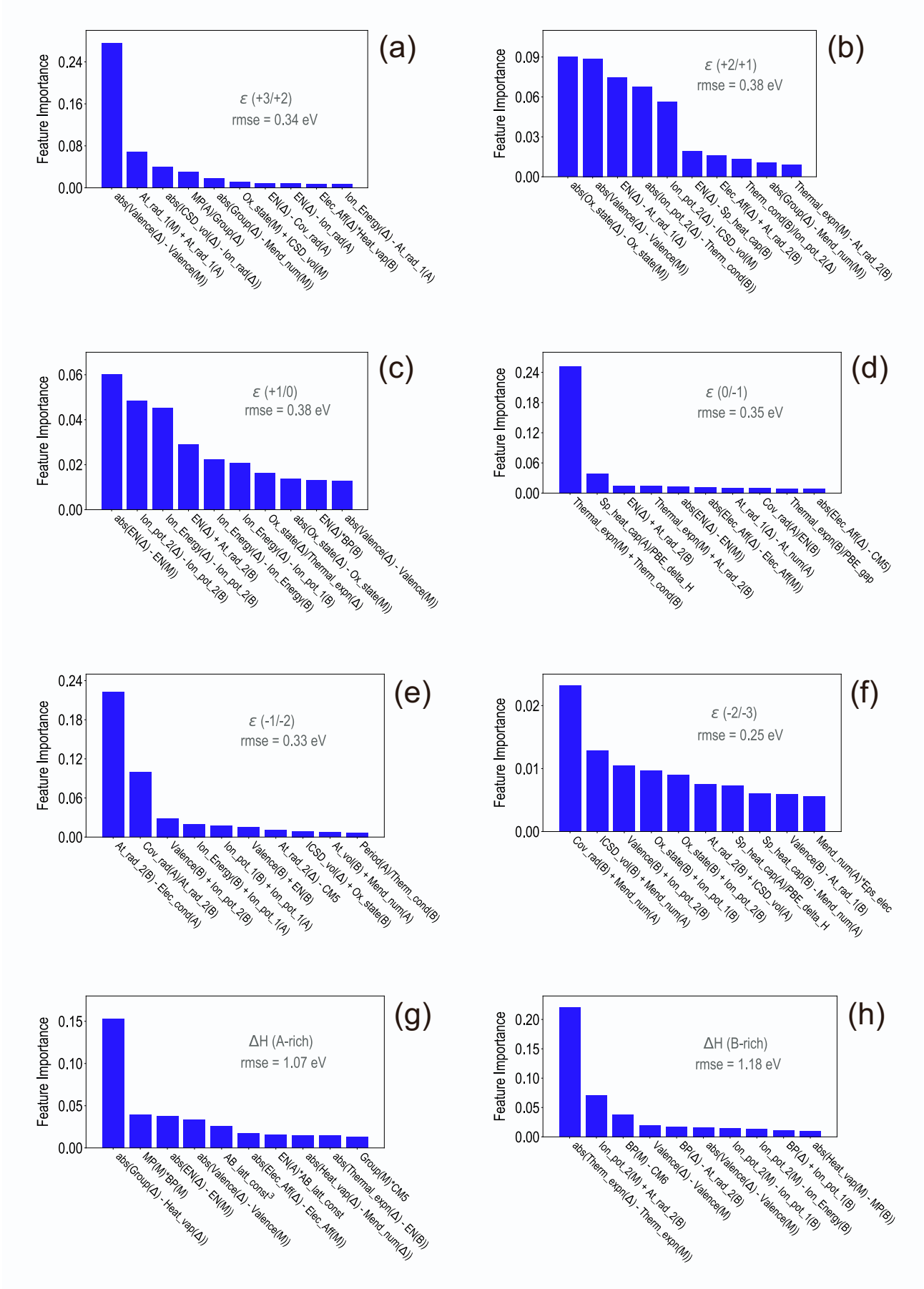

FIG. S10: Feature importance values for the 10 best SISSO-based compound features, obtained from the best random forest regression models for each property. Δ refers to the difference of the corresponding property (feature) between the impurity atom M and the element previously occupying the site it exists in (A, B or none).

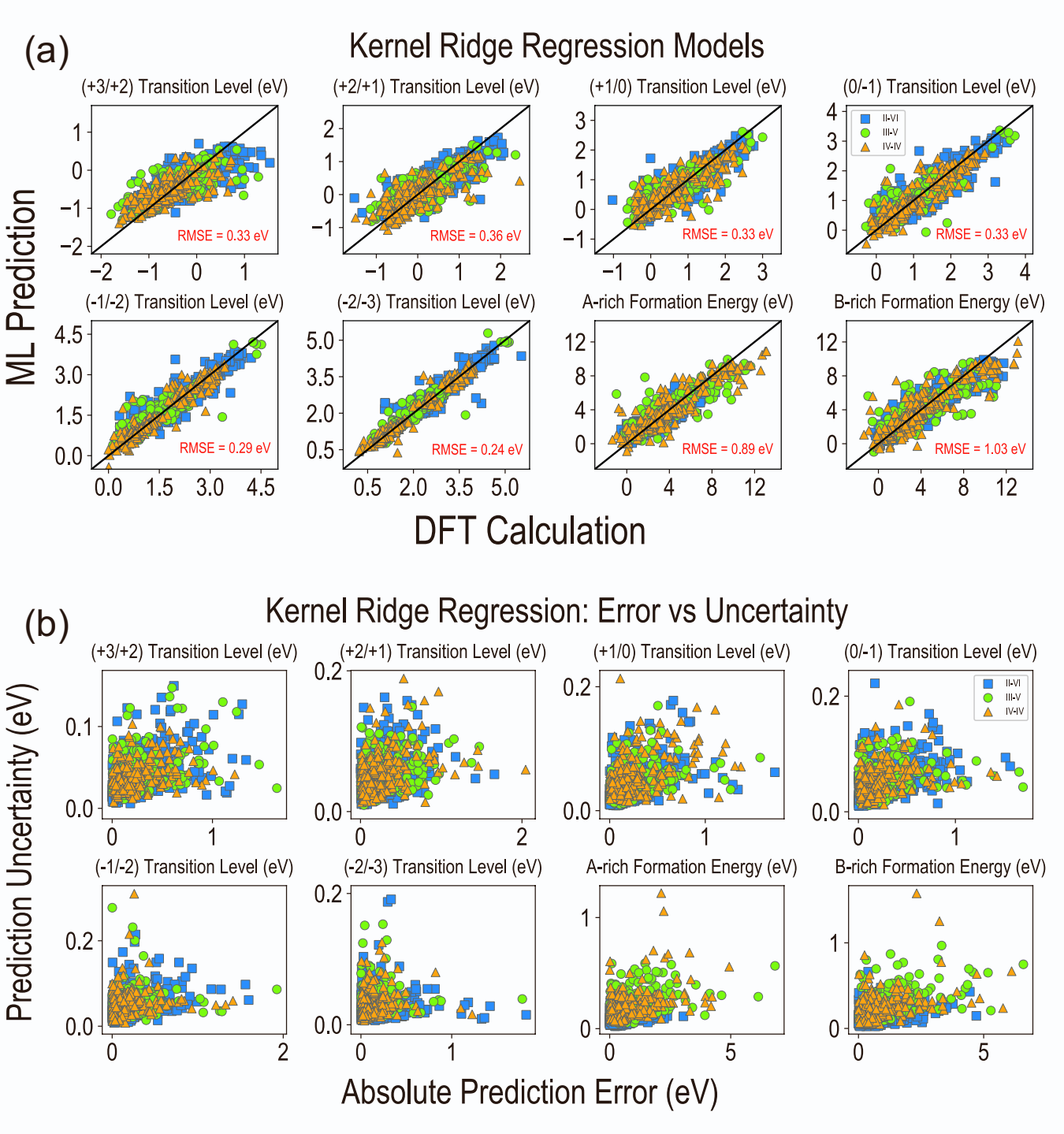

FIG. S11: Kernel Ridge Regression (KRR) results: (a) parity plots, and (b) prediction uncertainty as a function of absolute prediction error.

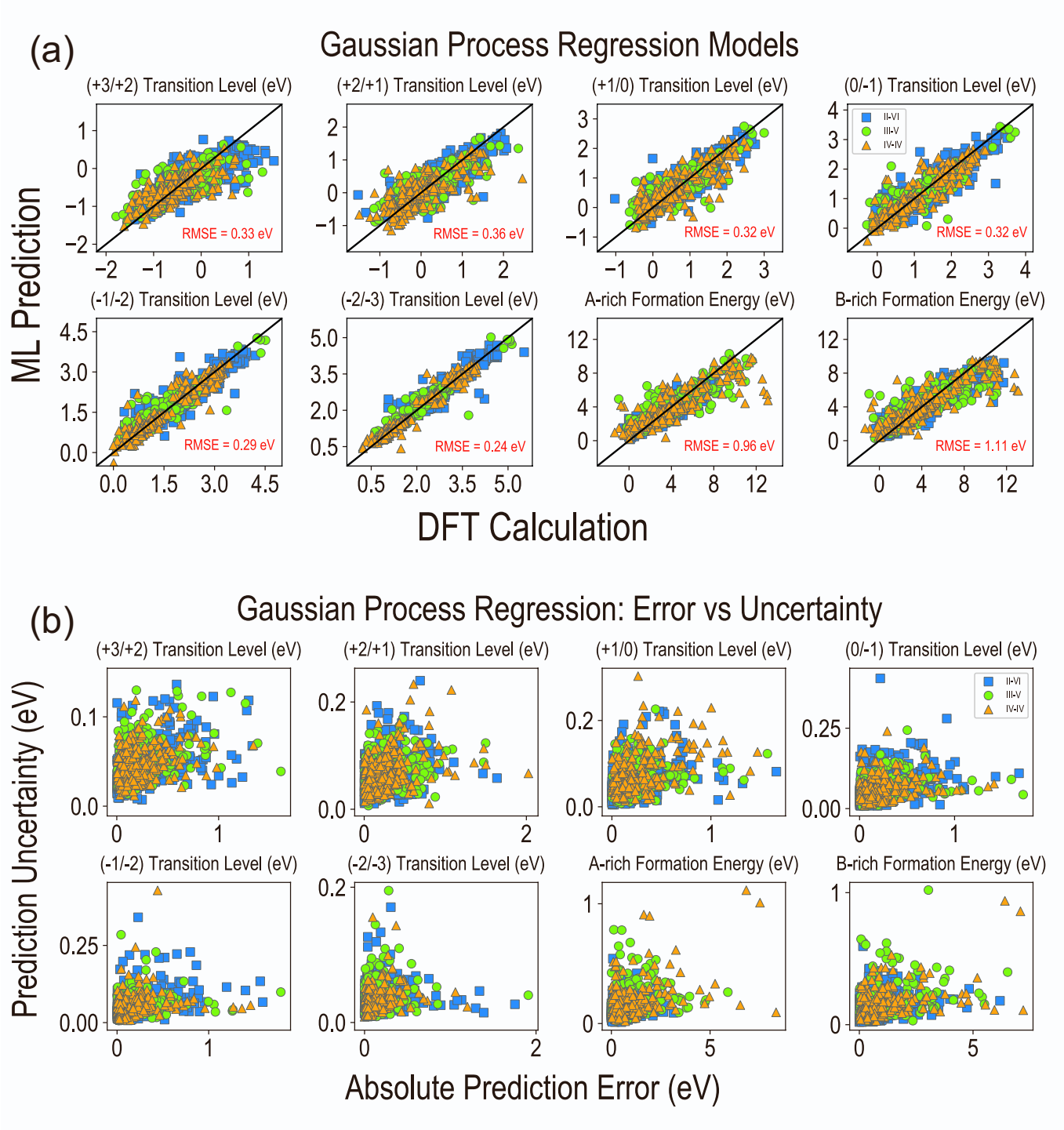

FIG. S12: Gaussian Process Regression (GPR) results: (a) parity plots, and (b) prediction uncertainty as a function of absolute prediction error.

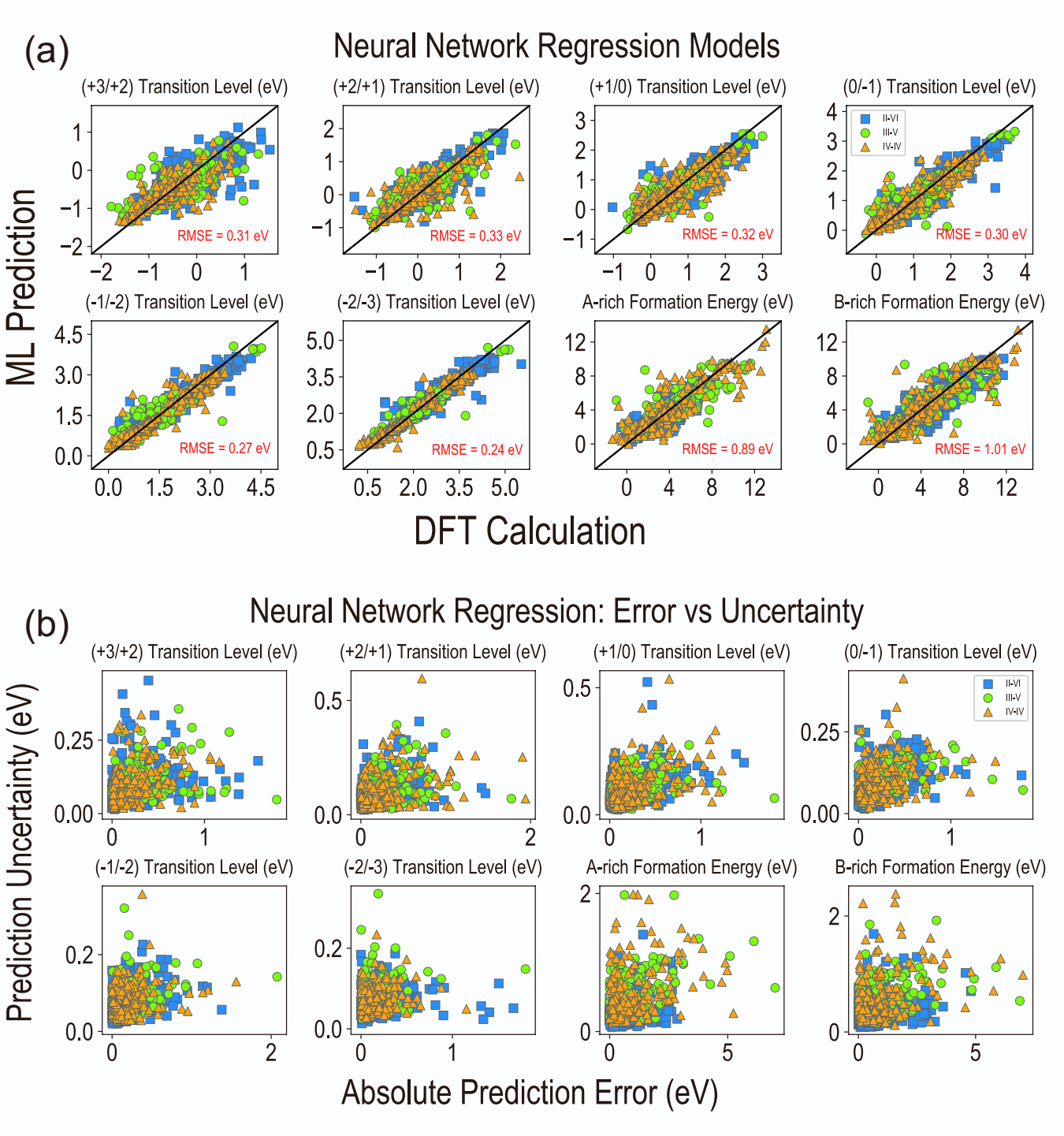

FIG. S13: Neural Network (NN) Regression results: (a) parity plots, and (b) prediction uncertainty as a function of absolute prediction error.

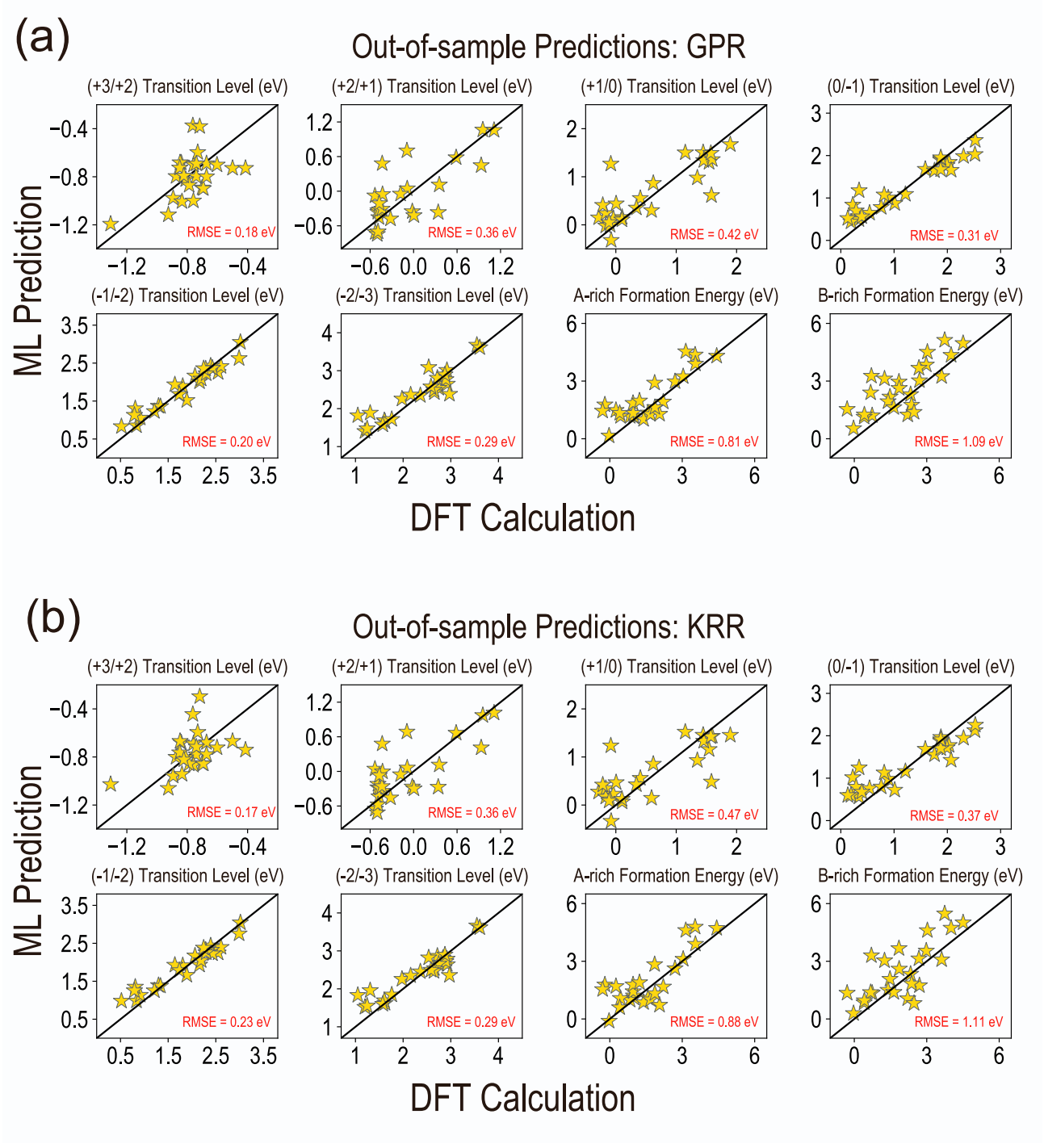

FIG. S14: GPR and KRR predictions compared to DFT results for 26 new, out-of-sample data points.

GPR Model, Entire Dataset	Test RMSE (Total)	Test RMSE (II-VI)	Test RMSE (III-V)	Test RMSE (IV-IV)	Test RMSE (Sub)	Test RMSE (Int)
ΔH (A-rich)	0.91±0.084	0.58±0.045	1.43±0.267	1.40±0.348	1.06±0.224	0.80±0.080
ΔH (B-rich)	1.00±0.120	0.71±0.092	1.51±0.382	1.45±0.272	1.15±0.261	0.90±0.068
(+3/+2)	0.35±0.015	0.34±0.034	0.37±0.072	0.33±0.047	0.34±0.027	0.35±0.022
(+2/+1)	0.38±0.030	0.35±0.033	0.38±0.062	0.42±0.060	0.44±0.047	0.34±0.045
(+1/0)	0.35±0.030	0.34±0.026	0.35±0.074	0.37±0.040	0.46±0.078	0.27±0.016
(0/-1)	0.32±0.013	0.31±0.023	0.34±0.066	0.31±0.033	0.42±0.037	0.25±0.033
(-1/-2)	0.27±0.014	0.26±0.017	0.28±0.070	0.29±0.064	0.31±0.028	0.25±0.020
(-2/-3)	0.23±0.013	0.23±0.037	0.23±0.055	0.21±0.034	0.21±0.033	0.24±0.017

GPR Model, Dataset Balanced by Semiconductor Type	Test RMSE (Total)	Test RMSE (II-VI)	Test RMSE (III-V)	Test RMSE (IV-IV)
ΔH (A-rich)	1.21±0.084	0.98±0.063	1.28±0.124	1.35±0.210
ΔH (B-rich)	1.38±0.099	1.29±0.023	1.38±0.061	1.41±0.210
(+3/+2)	0.36±0.030	0.38±0.032	0.37±0.049	0.32±0.024
(+2/+1)	0.41±0.020	0.42±0.004	0.40±0.030	0.41±0.051
(+1/0)	0.36±0.007	0.36±0.034	0.35±0.037	0.37±0.055
(0/-1)	0.36±0.029	0.36±0.034	0.37±0.070	0.35±0.040
(-1/-2)	0.32±0.018	0.34±0.042	0.33±0.030	0.28±0.060
(-2/-3)	0.25±0.027	0.28±0.034	0.25±0.055	0.20±0.024

GPR Model, Dataset Balanced by Defect Type	Test RMSE (Total)	Test RMSE (Sub)	Test RMSE (Int)
ΔH (A-rich)	0.97±0.096	1.05±0.225	0.85±0.130
ΔH (B-rich)	1.04±0.097	1.09±0.239	0.94±0.144
(+3/+2)	0.32±0.030	0.30±0.032	0.33±0.043
(+2/+1)	0.38±0.033	0.41±0.017	0.34±0.060
(+1/0)	0.36±0.038	0.43±0.038	0.26±0.050
(0/-1)	0.36±0.029	0.45±0.044	0.24±0.025
(-1/-2)	0.32±0.026	0.37±0.036	0.25±0.029
(-2/-3)	0.24±0.019	0.25±0.026	0.21±0.059

FIG. S15: RMSE comparison between GPR models trained on entire dataset and reduced balanced datasets.

SISSO/LASSO DATASET		
SISSO/LASSO manuscript predictions	train RMSE	test RMSE
dHA	0.632	1.07
dHB	0.711	1.18
(+3/+2)	0.222	0.34
(+2/+1)	0.262	0.38
(+1/0)	0.257	0.38
(0/-1)	0.217	0.35
(-1/-2)	0.23	0.33
(-2/-3)	0.19	0.25

NEW BALANCED DATASET		
SISSO/LASSO balanced type predictions	train RMSE	test RMSE
dHA	0.85	1.47
dHB	0.93	1.60
(+3/+2)	0.22	0.38
(+2/+1)	0.28	0.43
(+1/0)	0.28	0.41
(0/-1)	0.23	0.39
(-1/-2)	0.25	0.35
(-2/-3)	0.19	0.26

SISSO/LASSO balanced site predictions	train RMSE	test RMSE
dHA	0.66	1.14
dHB	0.73	1.23
(+3/+2)	0.20	0.34
(+2/+1)	0.27	0.41
(+1/0)	0.28	0.41
(0/-1)	0.23	0.38
(-1/-2)	0.25	0.35
(-2/-3)	0.19	0.26

FIG. S16: RMSE comparison between RFR models trained on entire dataset and reduced balanced datasets.